IDENTIFICACIÓN DE PATRONES DE DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 9 CON TÉCNICAS PREDICTIVAS DE MINERÍA DE DATOS

RICARDO MAURICIO ORTEGA MIPAZ
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
UNIVERSIDAD DE NARIÑO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
MAESTRÍA EN INVESTIGACIÓN DE OPERACIONES Y ESTÁ DÍSTICA
PASTO
2019

IDENTIFICACIÓN DE PATRONES DE DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 9 CON TÉCNICAS PREDICTIVAS DE MINERÍA DE DATOS

RICARDO MAURICIO ORTEGA MIPAZ

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARA OPTAR EL TITULO DE MAGISTER EN INVESTIGACIÓN DE OPERACIONES Y ESTÁ DÍSTICA

DIRECTOR:
Ph.D. RICARDO TIMARÁN PEREIRA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
MAESTRÍA EN INVESTIGACIÓN DE OPERACIONES Y ESTÁ DÍSTICA
PASTO
2019

TABLA DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1. ASPECTOS PRELIMINARES

1.1 Introducción

A partir del comienzo de los años 90, el Ministerio de Educación Nacional (MEN) comienza a investigar con mayor empeño sobre la aplicación de evaluaciones externas, fundamentadas en la medición de logros mediante las pruebas estructuradas por competencias.

Para el Ministerio de Educación Nacional es esencial que todos los colegios conozcan sus avances y puedan generar los cambios necesarios en pro del mejoramiento del sistema educativo. El Instituto Colombiano para la evaluación de la educación (ICFES), es el organismo encargado de realizar el seguimiento a la educación, mediante pruebas estandarizadas Saber 3°, 5° y 9°; los resultados obtenidos sirven como diagnóstico que permite interpretar y tomar medidas en aras de garantizar los respectivos ajustes, de acuerdo a las debilidades y/o fortalezas que presente los establecimientos educativos. (Ayala Garcia, 2015).

Los resultados conseguidos de las pruebas Saber 3°,5° y 9° permiten conocer el progreso de las instituciones, la eficiencia, el ambiente escolar, el desempeño de los colegios y secretarias de educación. Está evaluación es clave para hacer un seguimiento desde el punto de partida que la escolaridad inicia, promoviendo una evaluación formativa, evitando así deficientes resultados en el último año de la educación secundaria, que impiden el acceso a la universidad, (Fontecha Ariza, 2007), de acuerdo a lo mencionado anteriormente En este proyecto de investigación se propone recolectar información histórica de los resultados de las pruebas Saber 9° y crear una base de datos desde la cual se logre descubrir factores asociados al desempeño académico de los estudiantes de las instituciones educativas del país, que cursan grado noveno, y presentaron las pruebas Saber 9° entre los años 2014 a 2016, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, con técnicas descriptivas de Minería de Datos.

Se utilizará la metodología CRISP-DM, la guía más amplia empleada en el desarrollo de proyectos de Minería de Datos, que contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. El conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de

toma de decisiones de las instituciones gubernamentales y educativas que velan por la calidad de la educación en la República de Colombia.

1.2 Planteamiento del problema

El Ministerio de Educación Nacional tiene como propósito dentro de su política educativa, lograr una educación en alta calidad que garantice a los estudiantes del territorio nacional, oportunidades legítimas de progreso y prosperidad para ellos. (ICFES, Saber 5 y 9 Sintesis de resultados de Factores asociados., 2009). De ahí que, el Ministerio de Educación Nacional (MEN) en las últimas décadas ha construido innumerables programas de fortalecimiento a la educación, como son el programa de transformación de la calidad educativa (2011) todos aprender (2011,2014), Colombia la mejor educada en el 2005 entre otros, tácticas en el aula, capacitación a docentes y directivos, construcción de planes de mejoramiento, donde no solamente es participe la comunidad estudiantil sino la sociedad en general y además permitir contrarrestar las brechas de inequidad. (MEN, Ministerio de Educacion, 2019)

Con el objetivo de avanzar en el fortalecimiento de aprendizaje de los estudiantes, se definió que la evaluación censal y muestral de las pruebas SABER para los alumnos de 3°, 5° y 9° se realice periódicamente, buscando así monitorear los resultados del sistema nacional de educación a través de este tipo de evaluaciones y los resultados de estás hacen que los establecimientos educativos, secretarias de educación, Ministerio de Educación Nacional y sociedad en general conozcan cuáles son las fortalezas y debilidades que está presentando el sistema de educación nacional y a partir de ello apoyar la toma de decisiones entorno a acciones de mejoramiento por parte del Ministerio de Educación Nacional y los establecimientos educativos. (MEN, Colombia aprende la red del conocimient, 2018).

En este proyecto de investigación se busca identificar los patrones de desempeño académico en las pruebas saber 9 de las instituciones educativas que presentaron en el periodo comprendido entre los años 2014, 2015 y 2016, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES, aplicando técnicas de Minería de Datos.

1.3 Delimitación del problema

La presente investigación se enfocará en los resultados de pruebas SABER 9 desde el año 2014 hasta el año 2016 a nivel nacional, los cuales son obtenidos de las bases de datos del Instituto Colombiano para la Evaluación de la Educación (ICFES., Alineacion del examen SABER 11. Sistema Nacional de Evaluacion Estandarizada de la Educacion, Instituto Colombiano para la Evaluacion de la Educacion, 2013), con estos datos se buscará patrones de rendimiento académico mediante la técnica de Minería de Datos.

1.4 Viabilidad de la investigación

La investigación es viable porque se cuenta con documentos abiertos al público con acceso a las bases de datos de los resultados de las pruebas SABER 9 suministradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES).

1.5 Limitación de la investigación

Una de las limitaciones de la investigación es la calidad de los datos porque se tendrá en cuenta la evaluación censal que comprende instituciones educativas públicas y privadas, rurales y urbanas del país.

Otra limitación es la fidelidad y veracidad de la base de datos, porque al ser depurada se puede perder información relevante para el estudio de los factores asociados.

1.6 Justificación

Después de una revisión bibliográfica sobre educación y rendimiento académico se ha encontrado que existen documentos como (Velasquez, 2013) quien realiza un estudio entre los Estilos de Aprendizaje: Activo y Reflexivo de estudiantes de Grado Noveno del Nivel de Básica Secundaria, donde la relación que existe entre los estilos de aprendizaje (EA) y el rendimiento académico del año lectivo (RA) se evidencia una tendencia particular de aprendizaje reflexivo pero no respecto al rendimiento académico.

La diferencia de actitudes y aptitudes, también inciden en el desempeño de los estudiantes reflejado en parte, por su rendimiento académico puesto que, la continua variación de formas de relacionarse entre los repentinos cambios en asumir sobre quien recae la responsabilidad de aprender, dejan abierto el siguiente par de posibles incidencias: la primera en el sentido de una incoherencia que desfavorece el RA y la segunda como la adaptación inteligente que lo favorece. Otros estudios como (Gonzales, Caso, Diaz, & Lopez, 2012) es innegable en que el rendimiento académico esté asociado a diferentes factores internos de las instituciones educativas, como son: el número de estudiantes que tiene a cargo un docente, las instalaciones físicas de la institución, recursos tecnológicos con los que cuenta la institución, la ubicación geográfica entre otros; sin embargo existen factores externos a la institución como el contexto donde se encuentran los estudiantes, el grado de escolaridad de los padres, su nivel socioeconómico, el grado de afectividad de sus familiares, entre otros.

Por otra parte se encuentran otros estudios como (Fernandez, 2005) en los que entidades del está do muestran preocupación sobre el rendimiento de los estudiantes en las pruebas por su posicionamiento en relación a otros países. Esto hace que se asuman nuevas medidas para diseñar políticas educativas en búsqueda del mejoramiento de los aprendizajes en los estudiantes y por ende mejores resultados en las pruebas externas como las pruebas SABER.

Por esta razón este estudio es de gran importancia porque se encamina a identificar los patrones de desempeño académico por medio de la minería de datos en los estudiantes de grado 9 que presentaron las pruebas saber 9 desde al año 2014 hasta el año 2016, la cual puede servir de soporte al MEN, al ICFES y demás entidades está tales, como también a la comunidad educativa y a la comunidad en general, para la toma de decisiones pertinentes que contribuyan a mejorar la calidad de la educación en el país.

1.7 Objetivos

1.7.1Objetivo general

Identificar patrones de desempeño académico en las pruebas saber 9 de los estudiantes pertenecientes a instituciones educativas colombianas, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES en los periodos 2014

a 2016, a través de técnicas de Minería de Datos, que permitan generar conocimiento para soportar en las instituciones educativas y gubernamentales la toma de decisiones.

1.7.2 Objetivos específicos

- Seleccionar de las bases de datos del ICFES la información socioeconómica, académica e institucional de los estudiantes del país que presentaron las pruebas Saber 9° en los años 2014 a 2016.

- Construir un repositorio inicial de datos con los valores de los atributos socioeconómicos, académicos e institucionales de los estudiantes del país que presentaron las pruebas Saber 9°, a partir de las bases de datos del ICFES.

- Aplicar técnicas de limpieza y transformación de información al repositorio inicial de datos con el fin de obtener información limpia, correcta, consistente y discreta de las pruebas Saber 9°.

- Aplicar las técnicas predictivas de Minería de Datos más apropiadas, para identificar patrones de desempeño académico en las pruebas saber 9 utilizando una herramienta de Minería de Datos de software libre.

- Evaluar los patrones obtenidos para determinar el conocimiento descubierto acerca de los factores socioeconómicos, académicos e institucionales asociados al desempeño académico de los estudiantes que presentaron las pruebas Saber 9° a nivel nacional.

2. MARCO TEORICO

2.1 Antecedentes

Los estudios acerca de factores asociados pueden ayudar a comprender mejor el funcionamiento de los sistemas educativos y así identificar aspectos dentro de las escuelas que influyen positiva o negativamente sobre las oportunidades de aprendizaje de los alumnos. Los cuales constituyen un aporte que agrega valor a las evaluaciones estandarizadas (Ravela, 2006).

Así un tema determinante en la calidad de la educación en todo país es el rendimiento académico de sus estudiantes, el cual es entendido como el proceso de aprendizaje que depende de varias variables objetivas como son los sistemas de evaluación, las políticas educativas entre otras y variables subjetivas que se relacionan con los aspectos familiares, sociales, socioeconómicos, cognitivos entre otros (Erazo, 2012).

Un estudio que se destaca a nivel internacional es el titulado Rendimiento académico y factores asociados. Aportaciones de algunas evaluaciones a gran escala. Estudio realizado por (Gonzales C. J., 2012) en este trabajo se hace un análisis de factores asociados al rendimiento académico como parte de la aplicación de evaluaciones a gran escala donde se tienen en cuenta variables determinantes desde el nivel de sistema educativo, como son las pruebas internacionales PISA, las cuales tiene como propósito evaluar a los estudiantes de diferentes países en el desarrollo de competencias básicas que estos tienen.

En Colombia se han realizado varios estudios que buscan determinar los factores que influyen en el rendimiento académico de los estudiantes. Uno de ellos es el realizado por Gaviria y Barrientos (2001b), quienes analizaron los resultados de las pruebas de está do de 1999, donde encontraron que las características de la institución educativa influyen de manera significativa el rendimiento académico y que tienen mayor peso que las variables socioeconómicas; además, en este estudio también se encontró que el nivel educativo de los padres juega un papel importante en el desempeño académico. También evidenciaron, que existe una diferencia entre los resultados de instituciones oficiales y privadas. Estos hallazgos ponen en cuestión los resultados de (Coleman, y otros, 1996), quien concluyo que el rendimiento escolar en está dos unidos estaba influenciado por las características socioeconómicas de los estudiantes y no por el hecho de las variables asociadas a la institución educativa.

Otro estudio en este campo es el realizado (Villafañe, 2015), quien aplico técnicas de minería de datos al examen de está do saber 11 y encontró que el entorno socioeconómico del estudiante y su familia al igual que el nivel de escolaridad de los padres de familia influyen en el rendimiento académico; a mayor nivel socioeconómico y mayor nivel de escolaridad hay un mejor rendimiento académico.

Por otro lado tenemos la investigación contratada por el ICFES en el 2011 que fue realizada por Luis Jaime Piñeros quien hizo un estudio sobre eficacia escolar enfocada a los factores asociados a los resultados de los estudiantes en las Pruebas Saber 5° y 9° aplicadas en 2009. Que tuvo como propósito una mejor comprensión de aquellos aspectos de los contextos personales, familiares y escolares que inciden en los desempeños de los estudiantes en las pruebas, y aportar a la toma de decisiones de políticas orientadas a mejorar la calidad y la equidad de la educación en Colombia.

Otro estudio realizado por (Chica, Galvis, & Ramírez, 2010) Quienes en su investigación utilizaron el modelo Logit Ordenado Generalizado obteniendo como resultado la relevancia que tienen las variables socioeconómicas en el desempeño de áreas de matemáticas y lenguaje de las pruebas ICFES Saber 11º del semestre B de 2009.

Por otra parte, Gutiérrez, Y. (2015), analizó la relación que se presenta entre la estructura familiar de los estudiantes de grado 3° y 5° de primaria y el rendimiento académico en el área de Matemáticas en las Pruebas Saber del año 2013. Así los autores concluyen que el factor asociado que tiene mayor influencia sobre el desempeño académico de los estudiantes es el tamaño de la familia.

Ahora bien, la minería de datos en la educación es utilizada para comprender mejor el conocimiento que adquieren los estudiantes y así evaluar su desempeño académico y los entornos educativos en que aprenden, ayuda a identificar el éxito o el fracaso en las estrategias de enseñanza, aprendizaje y a generar un discernimiento más profundo del contexto educativo.

Las investigaciones llevadas a cabo en minería de datos en la educación se pueden realizar en el sistema educativo virtual y en el sistema educativo tradicional o presencial, aunque entre

estos dos sistemas se presentan diferencias por los contextos en que se desarrollan, es posible aplicar las diferentes técnicas de minería en ambos casos (Mr.Suhas G. Kulkarni, 2016).

La Minería de Datos en la educación no es un tema nuevo su estudio y aplicación ha sido muy relevante en los últimos años, se pueden utilizar las diferentes técnicas para describir y predecir distintos fenómenos dentro del campo de la educación (Timaran, Calderon, & Jimenez, 2013a) y (Timaran, Calderon, & Jimenez, 2013b). Un caso particular es predecir, con un porcentaje muy alto de confiabilidadad, la probabilidad de deserción de cualquier estudiante (Valero, 2009) (Valero,, Salvador,, & Garcia,, 2010).

Así las instituciones educativas pueden aprovechar las técnicas de Minería de Datos para realizar estudios acerca de las características de los estudiantes, de los métodos evaluativos y los diferentes factores asociados con lo cual se pueden descubrir procesos exitosos, detectar fraudes o inconsistencias que de otra manera permanecerían ocultos (Orea, Vargas, & Alonso, 2005).

En los últimos años, se ha incrementado el interés en utilizar la Minería de Datos en la educación, utilizando los datos proporcionados por las diferentes plataformas de educación centrándose en el desarrollo de nuevos métodos de descubrimiento para comprender mejor a los estudiantes y el entorno en el que aprenden. Los métodos empleados en la Minería de Datos en la educación suelen diferir de los métodos más generalistas, explotando explícitamente los múltiples niveles de jerarquía presentes en los datos (Jimenez, & Alvarez,, 2010)

Por otro lado, en la región de Nariño se han realizado algunos estudios que tienen como fin determinar los factores asociados al rendimiento académico, es el caso la investigación realizada por (Gomez & Jaramillo, 2017) quienes analizaron los resultados de las pruebas saber 5 de los periodos comprendidos 2014 al 2016, donde encontraron que la zona y el sector de la institución educativa son de gran incidencia en el desempeño académico, en dichas pruebas.

Además, (Acosta & Barahona, 2018), Realizaron un mercado de datos para almacenar información histórica de los resultados obtenidos por estudiantes que presentaron las pruebas saber 9 de los periodos comprendidos del 2014 al 2016 de las instituciones educativas de la

subregión de Obando en el departamento de Nariño, lo cual permite conjuntamente con la herramienta Pentaho realizar un análisis por múltiples variables y visualizar los resultados de manera gráfica lo cual facilita la toma de decisiones.

Por otro lado, (Timaran, Benavides, & Hidalgo, 2018) analizaron los factores asociados al rendimiento académico de las pruebas saber 11 encontrando que el estrato socioeconómico y el ingreso familiar de os estudiantes tiene gran influencia en el rendimiento académico de está s pruebas.

2.2 Conceptualización de pruebas saber 9

El Ministerio de Educación Nacional (MEN) delegado en propiciar y evaluar la educación colombiana a través de las Pruebas Saber (ICFES), han venido desarrollando la evaluación de registro para los bachilleres que terminan la educación secundaria y media mediante el Decreto 3156 del 26 de diciembre de 1968, con el fin de, ingresar a la educación de pregrado.

Las Pruebas Saber son una herramienta que permite evaluar la calidad de la educación en todos los colegios e Instituciones Educativas de Colombia. El propósito de aplicar está s pruebas es monitorear el rendimiento académico, entendido como el desarrollo de competencias básicas de los estudiantes y hacer seguimiento a la calidad educativa que se brinda en las instituciones educativas; también se consideró los avances de tiempo y espacio en cada periodo de la prueba y de sus programas, para luego, plantear acciones detalladas de mejoramiento. (ICFES, Documentacion de la prueba Saber 3, 5 y 9., 2018)

En 1994 estos exámenes fueron más importantes por la utilidad, al realizar una lista oficial de aspirantes, con el fin de dar a conocer la información obtenida a todas las instituciones educativas de Colombia, y comprobar sus fortalezas y falencias en dichas pruebas.

El diseño técnico de las pruebas SABER 3°, 5° y 9° para el período 2009 – 2021 está alineado con los Estándares Básicos de Competencias y se estructuró bajo está técnica, la que permitió definir y detallar los constructos que se evaluarán en lenguaje, matemáticas, ciencias naturales

y competencias ciudadanas, según corresponda en cada grado. Su diseño está alineado con los Estándares Básicos de Competencias establecidos por el Ministerio de Educación Nacional entendidos como referentes comunes a partir de los cuales es posible establecer qué tanto los estudiantes y el sistema educativo en su conjunto están cumpliendo con unas expectativas de calidad en términos de lo que saben y saben hacer.

El propósito de las Pruebas Saber (2009) para fortalecer la competencia es propiciar el mejoramiento de la calidad de la educación colombiana; "en ellas se valoran las competencias básicas de los estudiantes y se analizan los factores que inciden en su rendimiento escolar" (p.1).

Según el MEN, los resultados de está s evaluaciones permiten que los establecimientos educativos, las Secretarías de Educación y la sociedad educativa conozcan cuáles son las fortalezas y debilidades de los estudiantes y luego, pudieran definir planes de mejoramiento en sus respectivos ámbitos de intervención y todo esto alineado con los Estándares de competencias establecidos por el MEN.

A partir de esto, las pruebas permiten realizar un análisis sobre cambios frecuentes o probables que más influencian en el rendimiento académico de los estudiantes y de está forma, mejorar la calidad de la educación en el país, por tal motivo, expandió el proceso de evaluación por cada área e inició el estudio de los factores asociados al rendimiento escolar, para este fin se aplicó una consulta de prueba en el anuario del 2012 para recoger información y realizar un diagnóstico sobre el contexto de los estudiantes, sus familias y la institución educativa. Por lo tanto, el ICFES implementó un modelo llamado: contexto, insumos, procesos y productos (CIPP) esto permite verificar; cambios de un contexto particular para luego describir de forma detallada y precisa como éstas variables inciden en el rendimiento académico de los estudiantes. Figura 1: Modelo CIPP

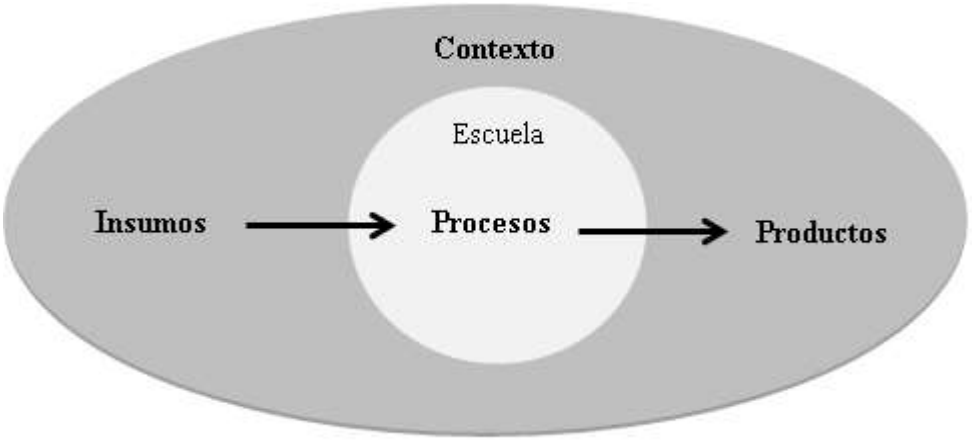

Figura 1: Modelo CIPP

En la categoría (ICFES, 2016), y dentro de está s variables se incorporan: la ubicación geográfica y aspectos económicos, políticos, culturales y sociales de la institución educativa, así como, la forma administrativa y su espacio académico. También aspectos relacionados con las características subjetivas de los estudiantes y su entorno según: su aspecto cultural, económico, condición sexual, perseverancia y las capacidades cognitivas del nivel académico de cada estudiante.

Del mismo modo, la materia prima hace mención a los elementos y recursos con los que cuenta cada Institución Educativa y colegios, estos recursos son muy importantes para determinar el desempeño académico de los estudiantes, se espera una buena utilización, por parte de la planta académica de cada Institución.

En está categoría, la infraestructura escolar es factor determinante para el desempeño académico de cada Institución en general, entre ellos están: el acceso a computadoras, la conexión a internet, pantallas de tv., para cada salón de clase y lo más importante; los servicios básicos. Por otra parte, el cronograma del año lectivo, las estrategias didácticas implementadas en los estudiantes al interior de los colegios; los cambios inesperados al trascurrir el año escolar, los estudiantes reprobados y la deserción escolar. (ICFES, Documentacion de la prueba Saber 3, 5 y 9., 2018)

En los sistemas o procesos se encuentran cambios que permiten caracterizar el ambiente escolar interno y externo, la parte administrativa de los directores y la labor cumplida de los docentes y de todo el plantel institucional. Así mismo, se considera la organización y el

desarrollo de las actividades que se implementan en las instituciones educativas para que los estudiantes se apropien de los conocimientos y logren desarrollar competencias establecidas desde el MEN.

Para concluir, la producción; es el resultado de las actividades empleadas en las entidades públicas y privadas como también de la misión y visión institucional, representado en el desarrollo intelectual, emocional, cultural y cívico que implica formar un ser crítico social.

El modelo CIPP representa direccionalidad de las relaciones que se espera encontrar a partir del planteamiento de opiniones en los factores asociados a la educación; es completo porque incluye todas las categorías de los factores asociados al aprendizaje y la influencia de está s en el proceso educativo, también es directo, porque presenta resultados que se pueden entender a partir de cualquier ámbito desde especialistas en investigación educativa, así como tomadores de decisiones, docentes y rectores a quienes también se des realizo un cuestionario que indagaba sobre los diversos factores que podían estar relacionados con el aprendizaje de estudiantes de 5 y 9. De igual forma, este modelo permite incluir cambios que son importantes para los procesos de enseñanza y desarrollo íntegro de los estudiantes.

2.2.1 Factores asociados

El estudio de los factores asociados a las pruebas Saber 3°, 5° y 9° publicado por el ICFES en el 2016, y dentro de cada una de las categorías incluidas en él; se especifica unos cambios que han sido conceptualizados y analizados a nivel internacional y pueden usarse de forma mecánica para responder a Planes de Mejoramiento Institucional PMI o adaptarse a cualquier situación en contextos ya sean regional, nacional e internacional y previamente en el marco de factores asociados se continua con el modelo CIPP. Figura 2

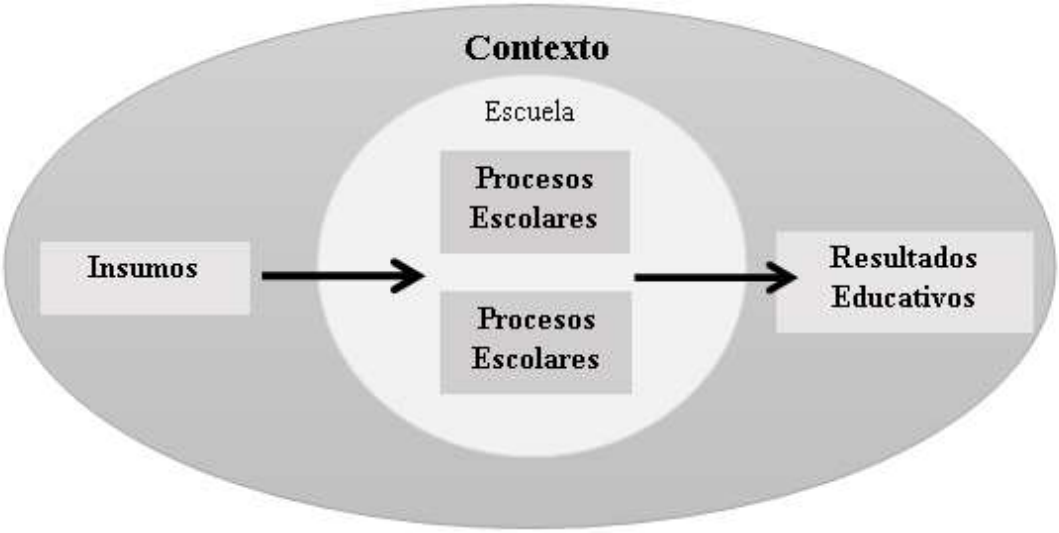

Figura 2. Marco de factores asociados Saber 3, 5, 9. 2016

2.2.2 Contexto

El propósito de los factores asociados es integrar los cambios de mayor relevancia al Plan de Mejoramiento Institucional (PMI) porque son los alcances que se consideran deseables, valiosos y necesarios, para la formación integral de los estudiantes, por lo tanto, se describen a continuación:

2.2.2.1 Características de los estudiantes

De acuerdo con Coleman (1966) y la UNESCO (2010), los aspectos sociodemográficos, de los estudiantes y sus familias son perspectivas fundamentales a la hora de analizar los resultados escolares. Estos factores influyen en el aprendizaje, pero no se pueden cambiar porque no dependen de las instituciones educativas. Esto evidencia que las variables que se examinan dentro de los factores asociados son:

- **Nivel socioeconómico:** La condición económica es variable y limitante frente al desempeño y desarrollo de los estudiantes en las Pruebas Saber, como lo afirma Célis, Jiménez & Jaramillo (2012), Icfes (2011). Por lo tanto, la disposición de las TICS (Tecnología de la Información y la Comunicación), la situación económica donde viven, el nivel de educación familiar, la explotación laboral, entre otras, son aspectos trascendentales para el rendimiento escolar.

- **Extra edad:** Para la ley general de educación un estudiante se encuentra en extra edad si supera en dos años la edad promedio del grado cursado; por tanto, la escolarización es obligatoria entre los 5 y 15 años de edad, desde primero a noveno grado (ICFES, 2016). Este cambio se relaciona con el grado que cursa el estudiante; permitiendo identificar aspectos como la repetición de año, la extra edad y deserción escolar.

- **Violencia intrafamiliar:** Según Chaux (2009), un entorno violento tiene influencias negativas en cuanto al rendimiento escolar y la formación ciudadana. Efectivamente, para determinar el tipo de violencia que se comete en el núcleo familiar y como esto afecta el rendimiento y desempeño académico de los estudiantes; se debe hacer una focalización y luego proceder a un diagnóstico, para concluir con qué frecuencia ocurren dichas situaciones y en qué tipo de violencia se está cometiendo.

- **Tipo de género:** Las brechas de desigualdad en el rendimiento académico de los niños y niñas, en ciertas ocasiones, se refleja en arquetipos de género que lamentablemente emergen desde la familia y la sociedad, por ende, decir que: la gran desventaja que se muestra en las niñas frente al aprendizaje de las matemáticas y a los niños en lectura, resulta ser una gran falacia, la Unesco, 2010 y 2016 a; Goldin et. al., 2006 y Niederle, et al., 2010.

- **Motivación:** La confianza y criterio que tenga cada estudiante sobre su desempeño, afecta de manera positiva o negativa su rendimiento académico; esto se evidencia en la participación y la asistencia a las clases, en el cumplimiento de tareas y evaluaciones. Desde el punto de vista de las Pruebas PISA, la motivación es reconocida como parte fundamental de los comportamientos o actitudes positivas frente al aprendizaje (OECD, 2003; OECD, 2010b; OECD, 2016a). Está variable es muy importante en el desempeño académico de los estudiantes es así como Ryan and Deci, (2000) proponen dos tipos de motivación: la intrínseca cuando se refiere a la motivación subjetiva y la extrínseca cuando es causada por algún incentivo en particular.

- **Explotación infantil:** Es una variable que afecta de manera negativa al rendimiento escolar ya que las condiciones físicas de los niños y niñas disminuyen, pues no están preparados para asumir tareas laborales pesadas. Pero se debe diferenciar entre el trabajo infantil que afecta negativamente el rendimiento escolar, y la colaboración en tareas del hogar, puesto que, esto muestra independencia y seguridad en sí mismos (Unesco 2015a; Unesco 2010).

- **Acompañamiento familiar:** La participación de los padres en el proceso de aprendizaje del estudiante, porque, beneficia el desarrollo cognitivo y emocional de los hijos. Genera actitudes de motivación frente a la formación académica, reflejadas en los resultados escolares, y ayuda a crear hábitos de estudio en familia (Pomerantz, Moorman & Litwack, 2007).

- **Necesidades inclusivas:** Es importante distinguir entre necesidades educativas cognitivas y físicas, puesto que en las primeras se debe buscar potenciar y alcanzar los logros mínimos planteados para cada grado, mientras que las necesidades físicas; se asocia más a la adecuación de espacios físicos que le permita la accesibilidad al establecimiento educativos del país, ya que incide significativamente al ingresar a un espacio académico (Unesco, 2005).

- **Cultura o etnia:** Las instituciones educativas están en la obligación y por decreto, de incluir y atender a los estudiantes pertenecientes a alguna etnia o raza en particular; teniendo en cuenta que, en algunos casos presentan condiciones económicas bajas y de difícil acceso a la educación. Por consiguiente, las Instituciones Educativas deben modificar su PEI para la flexibilización del aprendizaje y respetar las creencias étnicas de los estudiantes (Unesco, 2015b).

- **Auto-evaluación académica:** En la formación de la auto-evaluación académica se tiene en cuenta aspectos como la comparación ya sea personal, social, la percepción de los padres, profesores compañeros (Kurtz-Dostes y Scheneider, 1994).

- **Métodos de aprendizaje:** Las estrategias de aprendizaje en algunos casos, son las que facilitan la consecución del conocimiento y de nuevas habilidades.

Según el Diccionario de la Real Academia Española (2017), una estrategia "es el arte de dirigir operaciones militares o liderar un asunto. En matemáticas es un proceso regulable, conjunto de reglas que aseguran una decisión óptima en cada momento". Por lo tanto, existen diferentes estrategias de aprendizaje, pero se intensifican los estudios en las tradicionales y las cooperativas. Caracterizándose las primeras por el uso de la memorización como recurso de aprendizaje y las segundas por la creación colectiva del conocimiento (Weinstein, Husman, & Dierking, 2000). Esto evidencia que las estrategias tienen múltiples significados y todo depende del contexto en que se desempeñe.

2.1.1.2 Características de las escuelas

Son importantes los cambios y modificaciones que se realizan a los entes educativos y colegios. Entre las variables más notables se encuentran:

- **Ubicación de la institución educativa (urbana o rural):** Se han realizado estudios donde se evidencian que las instituciones se encuentran en la periferia del Departamento, existen posibilidades de encontrase con docentes empíricos, falta de materiales y medios tecnológicos, y se da una mayor deserción estudiantil, principalmente por la falta de recursos económicos de los padres que en muchas ocasiones emplean a sus hijos para realizar las labores del campo. Por consiguiente, en diversos establecimientos ubicados en las zonas rurales existe la metodología de escuela nueva o docente unitario para varios grados (ICFES, 2016).

Desde la perspectiva de Bowen & Bowen (1999) y la Unesco (2015b) está característica evidencia situaciones de conflicto en zonas periféricas del país y por ende procesos de marginación al interior de las instituciones educativas. Las repercusiones al respecto recaen principalmente en los estudiantes, frustrando su adecuado desarrollo académico y posteriormente los bajos desempeños en Pruebas Saber.

Por lo tanto, en la institución educativa urbana; el problema de la drogadicción, el alcoholismo y las pandillas, incrementando de manera acelerada los índices en edades tempranas, también generan deserción escolar, en consecuencia, esto afecta directa e indirectamente el rendimiento académico de los niños y niñas del país.

- **Administración de escuelas y colegios:** En la formación Educativa existen dos dependencias administrativas que son: públicas y privadas, la primera está supervisada por el gobierno de cada país, la segunda está regida por algún ente privado, en su gran mayoría, está orientada por la religión, la procedencia de los aportes o los llamados colegios en concesión donde se entrega la administración de un colegio que cuenta con infraestructura pública a uno privado. Según Chubb (2001) comenta que los entes privados exigen a todo el plantel educativo, un mejoramiento constante en la enseñanza de los estudiantes por la influencia económica que aportan los padres. Todo esto, no quiere decir que la educación privada arroje mejores resultados en el rendimiento académicos de los estudiantes (Ball, 2012).

- **Planta educativa**: El tamaño y el número de estudiantes en la Institución Educativa es muy importante, por una educación más personalizada; Crenshaw (2003) y Lamdin (1995) afirman que la cantidad de estudiantes atendidos en la institución educativa influye en el tipo de enseñanza y en el proceso de aprendizaje de los estudiantes, no obstante, Ajani & Akinyele (2014) y Stevenson (1996) se centran más en las condiciones físicas que se encuentra la institución educativa afirmando que si el docente tiene los medios necesarios (tecnológicos, técnicos, entre otros) por estudiante, produce los mejores resultados como un apoyo para el proceso de enseñanza y aprendizaje.

Otros factores que se tienen en cuenta para explicar este factor hacen referencia a la capacidad de la planta docente, por la cantidad de cursos y niveles, el tipo de educación: técnica, profesional, artística, científica, entre otros.

- **Nivel económico de las instituciones educativas**: La infraestructura y sus instalaciones cómodas, garantiza beneficios, motivación y posiblemente un mejor desarrollo

escolar, pues en algunos casos se da la heterogeneidad social y académica de los estudiantes, a nivel de aula y escuela, además el impacto en el desarrollo del plan de estudios, la calidad de las interacciones, programaciones culturales, recreacionales y la capacidad del docente para desarrollar distintas didácticas de enseñanza (Duflo, E. et al, 2011). Ciertas particularidades asociadas al nivel económico tienen que ver con el ingreso monetario, posesión de bienes, nivel de profesión de los padres.

2.2.3 Materiales y convenciones institucionales

A continuación, se presentan materiales y aspectos que proporcionan información válida para está investigación:

- **Antecedentes escolares:** Los conocimientos previos que adquieren los estudiantes fuera y dentro del contexto escolar, influyen de manera directa e indirecta en el rendimiento académico. Son factores asociados como la participación en la edad temprana o en contraposición, la repetición de grado (diferentes motivos) son antecedentes que traen consigo los estudiantes y se consideran favorables o desfavorables para alcanzar nuevos procesos de enseñanza y aprendizaje (ICFES, 2016).

- **Asistencia en educación infantil:** Es favorable para el desarrollo de la sociedad en general. En este sentido, él está do debe enfocarse en desarrollar programas infantiles que fortalezcan y apoyen este grado buscando igualdad, participación y calidad en el aprendizaje escolar. De igual forma, la vinculación a programas de educación preescolar ofrece características diferenciadoras de los párvulos; quienes asisten a educación preescolar formal llegan con mejor nivel de preparación para ajustarse a las exigencias escolares de la educación primaria y secundaria ofrecida en los distintos establecimientos del país.

- **Reprobar un grado:** Según Holmes (1989) y Roderick (1994) comentan que repetir un grado tiene efectos negativos a nivel académico y emocional a través del tiempo. Cuanto menor sea el grado de repetición mayores serán los efectos o consecuencias negativas

para el aprendizaje. Por lo tanto, reprobar un grado ha tenido un impacto positivo por parte de los estudiantes, porque el nivel económico de los estudiantes refleja un efecto negativo y costos en las familias del continente americano (Unesco, 2015a). En la actualidad Colombia es uno de los países de Latinoamérica con las tasas más bajas de reprobar grados en educación primaria; no obstante, al considerar las consecuencias que acarrea este fenómeno en la educación, resulta propicio revisar detalladamente el PEI., de los establecimientos educativos (Unesco, 2015b).

- **Tics:** Actualmente la sociedad está movida por las Tecnologías de la Información y la Comunicación (TIC´S) pues brindan múltiples posibilidades de acceder a información de manera más rápida y efectiva, razón por la cual incluir dentro de la práctica docente el uso de las TIC´S, reducidas al uso de computadoras, dispositivos electrónicos y el acceso a internet, es de vital importancia dentro de los procesos de enseñanza y aprendizaje. Uno de los aportes de las TIC´S es el aprendizaje colaborativo que puede darse; a nivel interinstitucional entre docentes y estudiantes y a nivel intrainstitucional facilitando la colaboración y el aprendizaje colectivo entre familias y centros educativos alrededor del mundo, y permitir el acceso a procesos formativos virtuales para todos los ciudadanos Gómez y Macedo (2010).

Según los resultados obtenidos en las pruebas PISA 2012 se evidencia pequeños efectos en los resultados, situación que se explica por la poca formación de docentes en cuanto a las capacidades de entendimiento y pensamiento frente a estrategias con el uso de las tecnologías y de estudiantes y sus inexistentes prácticas de pedagogía encaminadas hacia uso intensivo y adecuado de las TIC (OECD, 2015).

- **Infraestructura y materiales escolares:** La infraestructura como: computadores en excelente está do, biblioteca, aulas de clases amplias y completas, zonas verdes con espacios deportivos, baños adecuados para todo el plantel educativo, entre otros, es fundamental porque generan gran interés y motivación, pero no esencial en el rendimiento académico de los estudiantes de todos los grados (Unesco, 2015b).

- **Experiencia profesional y evaluación docente:** Un buen docente debe darle una buena utilización a los artefactos tecnológicos, como son: los libros virtuales, programas educativos de internet, entre otros. El uso y aprovechamiento que se les dé a estos recursos en el aula de clases durante el desarrollo de las sesiones pedagógicas define en gran medida los resultados esperados frente al logro educativo (ICFES, 2016).

De igual manera, es imprescindible que el docente con su amplia experiencia comprenda las fortalezas y debilidades de sus estudiantes, y aplique didácticas pedagógicas que le garanticen el aprendizaje de sus estudiantes, independientemente de sus características de origen (Barber & Mourshed, 2008).

- **Materiales gratuitos:** Los útiles escolares gratuitos son herramientas educativas esenciales para que los estudiantes se apropien de los conocimientos. Está s propuesta s las realizan las entidades educativas, ONG, especialmente en zonas de difícil acceso, estudiantes vulnerables y desplazados, entre otros. Así mismo, los útiles entregados a los establecimientos educativos no son los que realmente requiere la institución o en algunos casos son artefactos tecnológicos, pero sin una adecuada instrucción y lamentablemente terminan archivados (Reimers, DeShano dasílva, & Treviño, 2006).

- **Horas labórales en la institución:** Las horas o el tiempo que se dedica exclusivamente al aprendizaje, sin contar las interrupciones escolares, por recursos incumplidos por el Ministerio de Educación, Por otra parte, como lo comenta Murillo (2007) la inasistencia escolar, la indisciplina en el aula causada por algunos estudiantes, falta de energía en la utilización de los medios tecnológicos, son factores o inconvenientes no planeados que generan interrupciones y suspender los procesos de enseñanza y perdida del tiempo efectivo dedicado al aprendizaje, la inadecuada utilización de los artefactos tecnológicos, condiciones climáticas imprevistas, y en algunos casos, la improvisación y carencia de conocimientos por parte del docente.

2.2.4 Procesos institucionales

Los procesos, métodos y forma de enseñanza desarrollados en el salón de clases, se transforman, cambian y se modifican diariamente, según la perspectiva teórica que se practique en el aula de clases, y las relaciones cotidianas entre docentes y estudiantes promoviendo así el aprendizaje (ICFES, 2016), por tal razón, los procesos Institucionales se determinan en:

- **Metodología educativa**: Es menester la técnica o metodología del docente para interactuar con sus estudiantes de forma positiva, utilizando didácticas pedagógicas que estimulen y fomenten el aprendizaje por parte del docente, entre ellas están: la motivación, la identificación de saberes previos, actividades de práctica, complementación, aplicación de saberes, y evaluación de saberes. Por otra parte, el docente debe realizar métodos prácticos para que el estudiante logre captar la información pertinente, también debe realizar un seguimiento académico y disciplinar constante, que involucre un conjunto de operaciones intelectuales asociadas al conocimiento, garantizando en el estudiante aprendizaje. Y por eso, se llega a la capacidad del profesor para propagar y desarrollar el léxico en los estudiantes, a través de preguntas antes de iniciar una temática, indagar conocimientos previos de los estudiantes y el uso de un vocabulario técnico, entre otros (Pianta, Hamre, & Allen, 2012).

- **Convivencia institucional:** Es un clima escolar y agradable, pero sin perder el respeto a estudiantes, docentes y directivos; que se asocia con los logros académicos, propuestos por las normas y acuerdos que se instauran en cada institución educativa para propiciar un ambiente escolar acorde a las necesidades de cada estudiante y promover un aprendizaje significativo (OECD, 2016a). Por lo tanto, se procede a hacer una breve caracterización de cada una de ellas:

- **Normas y acuerdos:** La armonía y el respeto con normas fundamentales, que se lleven a cabo, son fundamentales en el clima escolar, fortalecen la enseñanza estudiantil, sin embargo, al instaurar despotismo y falta de respeto por parte del docente o los estudiantes y viceversa, esto produce indisciplina y pérdida de tiempo que podría dedicarse a los procesos del aprendizaje (OECD, 2016a).

- **Relaciones de comunicación escolar:** La habilidad y la actitud que tienen los docentes para resolver las dudas de sus estudiantes, la actitud de los estudiantes frente a los aprendizajes y el tipo de comunicación que se debe utilizar en un salón de clase son base fundamental para fortalecer los procesos educativos (Cohen, McCabe, Michelli & Pickeral, 2009)**.** Las relaciones de intercomunicación personal como: afectiva y positiva entre todo el plantel educativo, esto genera relaciones de confianza y optimismo.

- **Percepción sobre el aula y la institución educativa:** La práctica y experiencia del docente debe está r basada en aspectos como; claridad en los logros y metas planteados para el desarrollo de las clases de tal manera que los estudiantes comprendan el sentido de las estrategias didácticas, saber escuchar a los demás, el apoyo entre estudiantes, docentes y directivos, preguntar y propiciar relaciones sentimentales que permitan favorecer la permanencia de todo el plantel educativo y el desarrollo escolar. En la literatura se argumenta que existen creencias frente a las experiencias vividas por los alumnos en la escuela que afecta significativamente su desempeño académico y su comportamiento (Lester, Garofalo & Kroll, 1989; Wang & Holcombe, 2010).

- **Dirección institucional:** El liderazgo educacional es importante para facilitar la orientación de los objetivos planificados y encaminar los factores internos y externos en dirección a resultados institucionales en los procesos de enseñanza.

Los procesos escolares y las estrategias didácticas de aprendizaje deben ser tomados en cuenta por las personas a cargo de la institución educativa, esa dirección institucional hace parte del fortalecimiento de la visión institucional y las metas por cumplir, la rendición de cuentas por parte de los directivos y el liderazgo de los docentes (Leithwood, 1994).

- **Cooperativismo:** El trabajo en equipo permite fortalecer los procesos y proyectos educativos, como: la integración, la valoración y la capacitación en grupos sirve para mejorar las teorías y prácticas educativas (Elmore, 2010), (Unesco, 2010), (OECD, 2016a). El trabajo colaborativo entre colegas permite mejorar la enseñanza y el

desarrollo escolar; también, es importante interactuar las experiencias significativas en el salón de clase, porque conlleva a un progreso en sí mismo y de los estudiantes.

2.2.5 Resultados educativos

Los sistemas escolares, son primordiales para fomentar el desarrollo integral de los estudiantes, convirtiéndose en un gran apoyo básico para las competencias ciudadanas, el rendimiento, las disciplinas académicas y el bienestar del plantel educativo. Por esta razón es importante describir cada una de ellas como se hace a continuación.

- **Disciplinas formativas:** La importancia de la medición en los aprendizajes y estándares de las diferentes disciplinas académicas no solo hacen parte las distintas competencias básicas como el lenguaje y la aritmética, sino también las áreas que conforman el desarrollo integral del estudiante y su formación humanista. De igual forma como lo establece los estándares básicos de competencias y los derechos básicos de aprendizaje (DBA) es primordial tener en cuenta, un principio claro y sensato para realizar una crítica constructiva de una institución, a los estudiantes o a un sistema educativo en general (Ferrer, 2006).

- **Civismo:** La educación ciudadana de los valores sociales; su función es la participación democrática, la convivencia social, tolerancia en todos sus factores. De igual forma, todo esto corresponde a la formación académica y convivencia ciudadana de los estudiantes. Por lo tanto, los colegios e instituciones deben propagar la convivencia, el desarrollo de habilidades, comportamientos y respeto en relación a la vida en democracia, cumplir a cabalidad las leyes y normas pactadas, participación democrática, etc. (Espínola, 2005), (Westheimer & Kahne, 2004).

- **Bienestar particular:** El sentir emocional de cada individuo es muy importante para reconocer su motivación y que otras personas lo perciban, esto genera un bienestar r tanto particular como social, convirtiéndose en un factor esencial en el proceso de enseñanza de los estudiantes.

En algunas instituciones, los resultados académicos son más importantes que la motivación personal, en algunas ocasiones quedan apartadas; sin mayor importancia, debido a que está variable de tipo psicológico asociadas a aspectos emocionales no presentan características cuantitativas que permitan un cálculo, pero si son evaluadas; el bienestar individual se puede entender como la percepción de prosperidad subjetiva emocional y ligada a sus inteligencias múltiples importantes para su desarrollo académico, lo comenta Coleman D. (1995).

2.2.6 Las pruebas Saber

Las competencias son el conjunto de habilidades, destrezas y conocimientos que desarrolla una persona para comprender, transformar y participar de manera activa y crítica en el mundo que lo rodea. Las Pruebas Saber de 3, 5 y 9 es una evaluación estandarizada que realiza periódicamente el ICFES con el fin de medir la educación básica primaria y secundaria miden está s competencias genéricas; ellas pretenden indagar en los estudiantes cómo resuelven y cómo utilizan el saber adquirido en las diferentes áreas para luego evaluar su aprendizaje, utilizando un examen tipo 1 de selección múltiple con única respuesta.

El (MEN) infiere el propósito de la educación como el desarrollo de determinadas competencias y, en consecuencia, a está s como el objeto de la evaluación. Las diferentes competencias pueden desarrollarse a lo largo del proceso educativo. Existen dos competencias muy importantes a la hora de evaluar y son: las competencias genéricas, comprendidas como indispensables para el desempeño social y civil de todo ciudadano, independientemente de su oficio o profesión. Luego están las competencias (no genéricas) propias de oficios laborales particulares, que derivan de una capacitación específica (ICFES, 2013).

La planeación técnica de las Pruebas Saber 3°, 5°, 9°, Saber 11 y Saber Pro, está alineado con los Estándares Básicos de Competencias y se estructuró bajo está metodología. La función es; evaluar y calificar cuantitativamente el progreso o involución de los estudiantes después de aplicar la prueba en un determinado grado, por consiguiente, un alumno que presentó la prueba Saber 9 después de cuatro años aplica la prueba Saber 9° con los resultados obtenidos de las dos pruebas se puede comparar y hacer un análisis del proceso de aprendizaje que alcanzó en la primaria hasta llegar al bachillerato (ICFES, 2013).

Por lo tanto, en todas las pruebas de los exámenes SABER se puede apreciar una aproximación a la interpretación y comprensión de textos, a la argumentación por medio de explicaciones y justificaciones presentes en ellos y también a las posturas críticas y reflexivas que se pueden asumir frente a estos.

En otro orden de ideas las competencias genéricas son el fundamento esencial en la formación que reciben los estudiantes de la educación básica; puesto que le ayuda a resolver problemáticas de su diario vivir cotidiano, también le permite al individuo tener una visión diferente en cuanto a la participación de la producción laboral, en las relaciones propias en su entorno familiar, cuestiones sociales que aporten beneficios en pro de una convivencia en paz. Por lo tanto, las competencias ciudadanas deben ser excepcionalmente importantes porque forman estudiantes en valores éticos y morales, no genera una dependencia material, sino una dependencia por ser mejor persona, y todo esto debe desarrollarse a lo largo de todo el proceso educativo sin importar el ciclo de formación y deben de ser para todas las áreas, de tal manera que todas ellas contribuyan a la formación de está s competencias.

Rychen y Salganik (2006), comentan que una competencia es más que conocimientos y destrezas. Involucra la habilidad de enfrentar demandas complejas, apoyándose y movilizando recursos psicosociales (incluyendo destrezas y actitudes) en un contexto en particular.

El MEN enfocó las competencias genéricas como las competencias que desarrollan en el área de Lenguaje, Matemáticas y Ciudadanía. De tal manera que todo ciudadano debe está r en la capacidad de leer de manera comprensiva artículos, noticias, redactar cartas, correos electrónicos, sacar cuentas al momento de hacer sus compras, realizar un presupuesto y llevar su economía familiar, además, de ser un ciudadano que conozca sus derechos, deberes y está en capacidad de elegir un representante con la convicción de conocer y comprender sus propuestas.

Las competencias reconocen diversos grados de desempeño y de logros, expresados mediante indicadores, que permiten identificar los diferentes momentos o niveles de logro constituyendo una competencia determinada. Está s competencias son relativas y se transforman de acuerdo con las experiencias del individuo y con los procesos de aprendizaje (Torrado, M, 2000, p.121),

según las competencias y parafraseando a Rodríguez (2007), es decir, como un conjunto identificable y evaluable de conocimientos, actitudes, valores y habilidades relacionadas entre sí, que permiten desempeños satisfactorios en situaciones reales y en contextos específicos. En este sentido, una persona demuestra que es competente en la acción y no con la repetición de un saber determinado; en otras palabras, se evidencia la competencia cuando el conjunto de saberes se proyecta en acciones concretas que demandan su ejecución consciente; por lo tanto, las competencias se manifiestan en los desempeños que tiene el estudiante en situaciones específicas. El diseño de las pruebas fue desarrollado por especialistas en el área a desarrollar se tuvo en cuenta los Estándares Básicos de Competencias tomados como un referente común en cuanto a lo que deben saber y saber hacer los estudiantes durante su proceso educativo independientemente de su lugar de origen, modalidad educativa, condiciones culturales, sociales, económicas, religiosas, entre otras.

En el 2009 el ICFES diseñó las Pruebas Saber 3°, 5° y 9° de tal forma que garanticen evaluaciones censales para un periodo de doce años con el fin de visualizar en los resultados la evolución que ha tenido la educación en Colombia. Está s pruebas evalúan las competencias en Matemáticas, Lenguaje y Ciencias Naturales; cabe resaltar que éstas no permiten evaluar la totalidad de las competencias que deben adquirir los estudiantes durante su instancia en la escuela, pero si son un aliciente para que ellos continúen su formación profesional, laboral y social a lo largo de toda su vida (ICFES, 2011).

A continuación, se presenta la Tabla 1 que evidencia los procesos de las competencias genéricas teniendo en cuenta las competencias en Matemáticas, Lenguaje y Ciencias Naturales.

Tabla 1. Procesos de Competencias Genéricas

Lenguaje	Matemáticas	Ciencias Naturales	Competencias ciudadanas
-Lectura -Escritura	-Razonamiento y argumentación. -Comunicación, representación y modelación.	-Uso comprensivo del conocimiento científico. -Explicación de fenómenos. -Indagación.	-Aspectos cognitivos de las competencias ciudadanas. -Acciones y actitudes ciudadanas.

	-Planteamiento y resolución de problemas		

Fuente.: **Guía de lineamientos generales de Pruebas Saber 2009**

Podemos ver que en la Tabla 2 se evalúan los componentes que son los ejes verticales con los Estándares Básicos de Competencias permitiendo evidenciar las fortalezas y debilidades que tienen los estudiantes

Tabla 2. Componentes de Competencias Genéricas

Lenguaje	Matemáticas	Ciencias Naturales	Competencias ciudadanas
-Semántica -Sintaxis Pragmática	*Numérico – Variacional *Geométrico – Métrico *Aleatorio	-Entorno vivo -Entorno físico -Ciencia, tecnología y sociedad (CTS)	-Pensamiento ciudadano. Convivencia y paz. -Participación y responsabilidad democrática. -Pluralidad identidad y valoración de las diferencias

Fuente.: **Guía de lineamientos generales de Pruebas Saber 2009**

La metodología de las pruebas saber se basan en preguntas de selección múltiple con única respuesta, con cuatro opciones de respuesta A, B, C y D y sólo una de ellas es la correcta, el número de preguntas que deben contestar los estudiantes y en particular los de grado noveno son: 36 para Lenguaje, 48 en Matemáticas y 48 para ciencias Naturales.

Por otra parte, el Plan Nacional Decenal de Educación 2006-2016, y el ICFES han progresado en la alineación del Sistema Nacional de Evaluación Externa (SNEE), a través de la redistribución de los exámenes: en 2009 con un actual diseño de Saber 3°, 5° y 9°; en 2010 con el nuevo diseño de Saber Pro; en 2014 con variables en Saber 11°.

En las Pruebas Saber 5°, que se desarrolló en el año 2014, su dirección consistió en la introducción de competencias ciudadanas la cual presenta situaciones de análisis vivenciales que se relacionan con su propio contexto, es decir, sus amigos, el colegio, la familia, el aula, entre otros, utilizando un lenguaje más coloquial y situaciones menos complejas diferentes a las del grado 9° (ICFES, 2014). La formación o dirección posibilita examinar los resultados en distintos niveles educativos, está s Pruebas Saber evalúan las competencias de las diferentes áreas académicas, donde se incluyen las competencias ciudadanas (ICFES, 2014).

2.3 Descubrimiento de conocimiento en bases de datos

El Descubrimiento de Conocimiento en Bases de Datos, es un proceso importante que tal como lo afirman Fayard, Piatetsky-Shapior y Smith (1996) permite identificar patrones útiles, novedosos y correctos que facilitan el entendimiento del entorno, puesto que su versatilidad permite que se aplique en cualquier ámbito, obteniendo como resultado reglas o funciones; está s se obtienen del análisis, descubrimiento de datos iterativos y la minería de datos.

Teniendo en cuenta lo anterior, es posible establecer unas que permiten realizar este proceso de forma adecuada (ver ***Figura 3***):

a. Selección
b. Pre-procesamiento/limpieza
c. Transformación/reducción
d. Minería de datos (Data mining)
e. Interpretación /evaluación

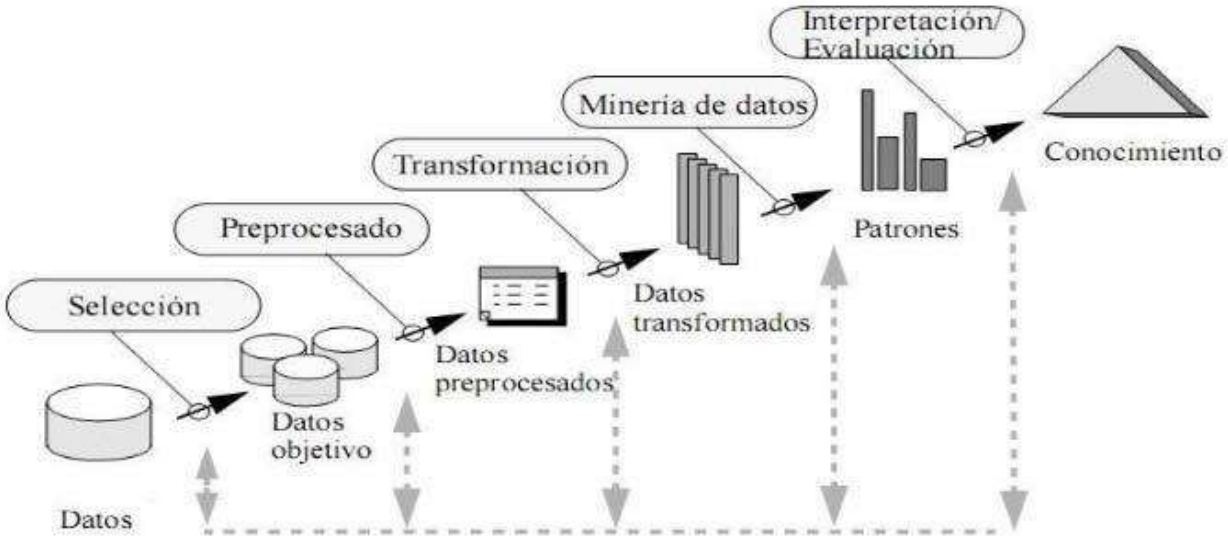

Figura 3.Etapas para la extracción del conocimiento. Fayyad et al.1996. Gómez Flechoso, 1998.

a. Etapa de selección: La identificación en los conocimientos relevante y prioritario donde se deben definir aquellas metas del proceso KDD, dentro del panorama que tiene el usuario final, se crea un conjunto de datos objetivos, fraccionando todo el conjunto

de datos que lo representan referente a la realización del proceso de descubrimiento, si bien la selección de los datos varía de acuerdo con los objetivos del negocio.

b. **Etapa de pre-procesamiento o de limpieza (Data cleaning):** se tendrá en cuenta la calidad de los datos, aplicando aquellas operaciones fundamentales en la remoción de datos ruidosos, las estrategias utilizadas serán para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas está dísticas para su sustitución. la etapa es de suma importancia ya que abarca la interacción con el usuario o analista. Incluso los datos ruidosos (noisy data) son valores que están elocuentemente distanciado del rango de valores esperados; esto se deben principalmente a errores que son cometidos por los humanos, o a esos cambios en el sistema, también se da por la información no disponible a tiempo o por las fuentes heterogéneas de datos. Los datos extraños (empty) son aquellos que no les corresponde un valor en el mundo real y en cambio los missing son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas administradores de bases de datos relacionales. Por consiguiente, en el proceso de limpieza, estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como la media, la moda, mínimo, máximo, desviación estándar para reemplazarlos.

c. **Etapa de transformación o reducción de datos**, es reconocida por esos datos que tienen una meta en el proceso, para ello se utiliza los métodos de reducción de dimensiones o de transformación, puesto que disminuye el número efectivo de variables bajo consideración o para encontrar representaciones inalterables de los datos (Fayyad et al., 1996). también este método se emplea para simplificar la tabla de una base de datos horizontalmente o verticalmente las cuales consisten:

- **La reducción horizontal:** hace referencia a la eliminación de tuplas (una secuencia ordenada) idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos.

- **La reducción vertical:** comprende a la eliminación de propiedades que son insignificantes o repetitivas con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (edades y fechas). Se utilizan técnicas de reducción de las añadiduras, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras (Han & Kamber, 2001).

d. Etapa de minería de datos: su objetivo se centra en la indagación y descubrimiento de patrones insospechados y de interés, aplicando esas tareas de descubrimiento tales como clasificación (Quinlan, 1986), (Wang, Iyer, & Vitter, 1998), clustering (Ng & Han, 1994), (Zhang, Ramakrishnan, & Livny, 1996), patrones secuenciales (Agrawal & Srikant, 1995) y asociaciones (Agrawal & Srikant, 1994), (Srikant & Agrawal, 1996), entre otras. Las técnicas que aplica la minería de datos consiste en la creación de modelos que son predictivos o descriptivos, dicho modelos que predicen persiguen entablar los valores futuros o desconocidos de las variables de interés, que se denomina variable de objetivo, dependientes o clases, donde son usadas otras variables denominadas independientes o predictivas, además sirven para determinar posibles nuevos o futuros resultados, si pierden o no en función de su zona de procedencia, facultad, estrato, género, edad y de más determinantes . Entre las tareas predictivas están: clasificación y regresión.

En los modelos representativos se identifican patrones que explican o resumen los datos. Su función es explorar todo el potencial de las propiedades de los datos examinados, no para predecir nuevos datos, sin embargo, en medio de las tareas descriptivas se cuentan: reglas de asociación, patrones secuenciales, clustering y correlaciones. Por tanto, la escogencia de un algoritmo de minería de datos incluye: la selección de los métodos a aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) a utilizar.

e. Etapa de interpretación o evaluación de datos. se pretende encontrar es la interpretación de los patrones descubiertos, añadiendo que, potencialmente se desatan las anteriores etapas para las posteriores reiteraciones, en donde aquella etapa, incluye

> la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones, cuyos términos deben ser útiles para que el usuario logre entenderlos. Además, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas, como también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. (Timaran, Calderon, & Jimenez, 2013b)

2.3.1 Tareas de minería de datos

La minería de datos comprende diversas tareas que pueden ser resueltas mediante el uso de algoritmos; dentro de estas tareas tenemos las tareas predictivas las cuales tratan problemas y tareas en los que hay que predecir uno o más valores para uno o más ejemplos. Dependiendo de cómo sea la correspondencia entre los ejemplos y los valores de salida y la presentación de los ejemplos podemos definir varias tareas predictivas, dentro de estás se encuentra la clasificación, clasificación basada en arboles de decisión, la regresión.

También se tienen las tareas descriptivas las cuales buscan describir los datos existentes como son: la segmentación o clustering, Asociación

2.3.2 Clasificación

Es un proceso que permite descubrir propiedades comunes en un grupo de objetos dentro de una base de datos, los cuales se cuentan de acuerdo al modelo de clasificación escogido, razón por la cual la clasificación se realiza en dos pasos; la primera consiste en crear un conjunto de tuplas extraídas de la base de datos, dichas tuplas se caracterizan por contener información determinada de la base de datos, los cuales se denominan atributo clase; a su vez el conjunto de estos recibe el nombre de conjunto de entrenamiento, cuya función es escoger aleatoriamente un número de tuplas extraídas de la base de datos. Cada tupla perteneciente al conjunto de entrenamiento se denomina ejemplo de entrenamiento. Una vez agotada está etapa se prosigue a la segunda la cual consiste en evaluar la exactitud del modelo usado mediante la utilización de un conjunto de prueba igualmente elegido de forma aleatoria y que es independiente del conjunto de entrenamiento. Dentro de este conjunto de prueba las tuplas contenidas reciben el

nombre de ejemplo de prueba, una vez comprobada la exactitud del conjunto de entrenamiento se obtiene el porcentaje de ejemplos correctamente clasificados. La importancia de la exactitud del modelo de clasificación recae en la posibilidad de usar la mejor tupla para las próximas tuplas futuras. Tomando en cuenta lo anterior, a continuación, se hace alusión a los diferentes métodos de clasificación: rough sets, árboles de decisión, redes neuronales, Bayes, algoritmos genéticos entre otros.

El modelo de clasificación basado en árboles de decisión. Este modelo de clasificación es el más utilizado probablemente por su simplicidad facilidad para entender (Han & Kamber, 2001), (Sattler & Dunemann, 2001), cuyo origen se remonta a los estudios de Aprendizaje de Máquina, permite construir arboles de decisión a partir de un conjunto de entrenamiento conformado por un conjunto de casos en contra posición con un conjunto de prueba. La calidad del árbol depende de la precisión de la clasificación y el tamaño del árbol, razón por la cual primero escoge un subconjunto del conjunto de entrenamiento y forma un árbol de decisión. Si el árbol no da la respuesta correcta para todos los objetos del conjunto prueba, una selección de excepciones se adiciona al conjunto de entrenamiento y el proceso se repite hasta que se encuentra el conjunto de decisiones correctas, lo que obtiene como resultado un árbol en el cual cada hoja lleva un nombre de la clase y cada nodo interior especifica un atributo con una rama correspondiente a cada posible valor del atributo.

El árbol de decisión se construye de la siguiente forma:

- Calcular la entropía que puede reducir cada atributo.
- Ordenar los atributos de mayor a menor capacidad de reducción de entropía.
- Construir el árbol de decisión siguiendo la lista ordenada de atributos

Entre los algoritmos de clasificación para árboles de decisión se cuentan ID-3 (Quinlan, 1986), C4.5 (Quinlan, 1993), SPRINT (Shafer et al., 1996), SLIQ (Metha et al., 1996) y J48 (Hall, Frank & Witten, 2011). La idea básica de estos algoritmos es la de construir los árboles de decisión donde, cada nodo no terminal está etiquetado con un atributo; cada rama que sale de un nodo está etiquetada con un valor de ese atributo; cada nodo terminal está etiquetado con

un conjunto de casos, cada uno de los cuales satisface todos los valores de atributos que etiquetan el camino desde ese nodo al nodo inicial.

La aplicación de un atributo determinado como criterio de selección clasifica los casos en distintos conjuntos. Para la construcción del árbol de decisión más simple y que sea consistente con el conjunto de entrenamiento, se requiere ordenar los atributos relevantes, desde la raíz a los nodos terminales, de mayor a menor poder de clasificación. Puesto que el poder de clasificación del atributo determinado, permite generar particiones del conjunto de entrenamiento, de tal forma que se ajuste en un grado dado a las distintas clases posibles, introduciendo de está forma un orden en dicho conjunto. También el poder de clasificación facilita la reducción de la incertidumbre o entropía (grado de desorden de un sistema). Está métrica se denomina ganancia de información. El atributo con la más alta ganancia de información se escoge como el atributo que forme un nodo en el árbol (Quinlan, 1993) (Agrawal et al., 1992)

La ganancia de información obtenida por el particionamiento del conjunto T, de acuerdo con el atributo A se define como:

$$Gain(T, A) = I(T) - E(A)$$

Donde $I(T)$ es la entropía del conjunto T, compuesto de s ejemplos y m distintas clases C_i $(i = 1, m)$ y se calcula: $I(T) = -\sum Pi\ log_2\ (Pi)$ don de $Pi = s_i/s$ es la probabilidad que un ejemplo cualquiera pertenezca a la clase *Ci* y *Si* es el numero de ejemplos de *T* de la clase *Ci*.

$E(A)$ es la estropia del conjunto *T* si está particionado por los n diferentes valores del atributo A en n subconjuntos {S1, S2, …,Sn}, donde Sj contiene esos ejemplos de *T* que tienen el valor aj en A y S_{ij} el numero de ejemplos de la clase *Ci* en el subconjunto S_j. $E(A)$ se calcula:

$E(A) = \sum \frac{S_{ij}}{S} * I(S_{ij})$ donde S_{ij} ejemplos de la clase *Ci* en el subconjunto *Sj*

$I(S_{ij}) = -\sum Pij\ log_2\ (Pij)$ donde $P_{ij} = S_{ij}//S_j/$ es la probabilidad de que un ejemplo de *Sj* pertenezca a la clase *Ci.* En otras palabras $Gain(T, A)$, es la reducción esperada de la entropía causada por el particionamineto de *T* de acuerdo al atributo *A.*

Finalmente, las reglas de clasificación se obtienen recorriendo cada rama del árbol desde la raíz hasta el nodo terminal. El antecedente de la regla es la conjunción de los pares recogidos en cada nodo y el consecuente es el nodo terminal.

Algoritmo Part, Este algoritmo combina los métodos de C45 y Ripper y no necesita realizar una optimización global para producir un conjunto de reglas precisas, la simplicidad es su principal ventaja dado que adopta la estrategia de separar y conquistar, ya que construye una regla elimina las instancias que cubre y continua creando reglas de forma recursiva para las instancias restantes hasta que no queda ninguna instancia. Difiere del enfoque estándar en la forma de crear cada regla, en esencia para hacer una sola regla para la cual se construye un árbol de decisión podado con el conjunto actual de instancias, la hoja con la mayor cobertura se convierte en una regla y el árbol se descarta.

Este proceso evita la que se generalice en forma apresurada, una vez que se conocen las implicaciones (es decir todos los subárboles que han sido ampliados); la posibilidad de construir repetidamente arboles de decisión y descartarlos no es tan extraño dado que se usa el árbol podado para obtener una regla, al utilizar la metodología de separar y conquistar junto con la decisión de árboles le agrega flexibilidad y rapidez a la misma. (Fibe Frank, 1998)

Algoritmo Zero R, este algoritmo de clasificación es el más simple que existe y depende solo del objetivo ignorando los predictores. Este clasificador simplemente predice sobre la clase o categoría principal, es útil para determinar una base sobre las cual se miden los demás métodos de clasificación. Este método utiliza la media para las variables numéricas o la moda para variables nominales en la variable de salida para asignar un único valor de probabilidad para todos los individuos. (Science, 2019)

La Segmentación o clustering, entendida como aquella clasificación de objetos físicos o abstractos sin supervisión (Chen, Han & Yu, 1996). Por tanto, para analizar el clustering se parte de la construcción de particiones significativas de un gran conjunto de objetos basado en la metodología "divide y conquista", motivo por el cual este método consiste en fragmentar en pequeños segmentos de propiedades similares, una base de datos y de esa forma simplificar el diseño y la implementación.

La principal característica de los clúster es su elevada homogeneidad y a su vez la heterogeneidad; por homogeneidad se debe comprender primero qué homogeneidad, es entendida como los registros próximos entre los fragmentos con respecto a la distancia entre el centro del segmento. En segundo lugar, la heterogeneidad se manifiesta como aquellos registros en diferentes segmentos que no son similares de acuerdo a una medida de identidad (Cabena,

1998). Es posible afirmar que el Clustering es el aprendizaje de conocimiento no supervisado toda vez que mediante el uso de algoritmos tales como K-Means (Han & Kamber, 2001), CLARANS (Clustering Large Applications based upon Randomized Search) (Ng & Han, 1994), y BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang, Ramakrishnan & Livny, 1996). Los cuales permiten encontrar subdivisiones homogéneas de datos, donde pueden tener aplicaciones prácticas de acuerdo al campo que se aplique este método.

2.3.3 Clasificación basada en asociación

La canasta de mercado propuesta por Agrawal et al. (1992) permite descubrir patrones en forma de reglas, los cuales representan hechos que ocurren de forma periódica en conjunto.

Ahora bien, se resalta que es una problemática, la cual conlleva a un conjunto de ítems que se encuentran subdivididos, cuya labor es encontrar una relación, que permita descubrir reglas asociativas a partir de la canasta, las cuales brindan soporte y confianza.

En las transacciones efectuadas entre ítems comprados, no se considera a cada ítem en forma individual si no como una variable binaria. Formalmente,

Sea $I = \{i1, i2, \dots im\}$ un conjunto de literales, llamado ítems. Sea D un conjunto de transacciones, donde cada transacción *T* es un conjunto de ítems tal que $T \subseteq I$. Cada transacción se asocia con un identificador llamado *TID.* Sea X un conjunto de ítems. Se dice que una transacción T contiene a X si y solo si $X \subseteq T$. Una regla de asociación es una implicación de la forma $X \rightarrow Y$ donde X, Y son conjuntos de ítems tal que $X \subset I,\ Y \subset I\ \ y\, X \cap Y = \Phi$. El significado intuitivo de tal regla es que las transacciones de la base de datos que contienen X tienden a contener Y. la regla $X \rightarrow Y$ se cumple en el conjunto de transacciones D con una confianza c si el c% de las transacciones en D que contienen X también contienen Y. la regla $X \rightarrow Y$ tiene un soporte s en el conjunto de transacciones D si el s% de las transacciones en D contienen $X \cup Y$.

Con el objetivo de conseguir reglas fuertes que brinden confianza y soporte entendido este como la frecuencia de ocurrencia de los patrones en la regla. Es encuentra que existe un obstáculo que no permite encontrar fácilmente reglas de asociación, sin embargo, existen algunos pasos a seguir para dar con ellas:

- Descubrir los ítemsets frecuentes, i.e., el conjunto de ítems que tienen el soporte de transacciones por encima de un predeterminado soporte s mínimo.
- Usar los ítemsets frecuentes para generar las reglas de asociación para la base de datos.

A continuación, después de que los itemsets frecuentes son identificados, las correspondientes reglas de asociación se pueden derivar de una manera directa. Un ejemplo de una regla de asociación es "el 30% de las transacciones que contienen cerveza también contienen pañales; el 2% de todas las transacciones contienen a ambos ítems" (Agrawal et al., 1996). Aquí el 30% es la confianza de la regla y el 2%, el soporte de la regla.

Es posible clasificar las reglas de asociación en unidimensionales y multidimensionales Según Han y Kamber (2001). Una regla de asociación es unidimensional, si los ítems o atributos de la regla hacen referencia a un solo predicado o dimensión. Ahora bien la regla de asociación es multidimensional, si los ítems o atributos de la regla hacen referencia a dos o más criterios o dimensiones.

Patrones secuenciales: su principal objetivo es encontrar un conjunto de secuencias cronológicas, llamado data secuencia, la cual consiste en una lista de transacciones que agrupa un conjunto de ítems concatenados. Este método se utiliza en el análisis de la canasta de mercado y comportamientos de compra. La problemática que conlleva se centra en la secuencia de los patrones los cuales deben cumplir con el soporte mínimo estipulado por los usuarios, donde dicho soporte se entiende como el porcentaje de la data de secuencia que debe contener el patrón. Sin embargo, todos los ítems en un elemento de un patrón secuencial deben está r presentes en una transacción simple para que la data-secuencia soporte al patrón (Agrawal & Srikant, 1995). Los patrones secuenciales, en el dominio de la medicina, por ejemplo, se pueden utilizar para ayudar a identificar síntomas y enfermedades que preceden a otras enfermedades.

2.3.4 Metodología CRISP-DM

Se utiliza para proyectos como son la minería de datos, debido a que permite comprender está tecnología y extraer ideas. CRISP-DM, consiste en un conjunto de tareas que están

organizadas en cuatro niveles de abstracción: fases, tareas generales, tareas especializadas e instancias de proceso. Cada nivel sigue una jerarquía de tareas que van de lo general a lo específico (Chapman et al. 2000), está figura se asemeja a la representación de un ciclo de vida en un proyecto de minería de datos y es muy versátil pues pese a la jerarquización de actividades y los pasos a seguir, no existe inconveniente en omitir ese orden debido que, el avance depende del resultado de cada fase. (Larose y Larose, 2014). CRISP-DM está compuesta por seis fases: análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación (Gallardo, 2009) como se aprecia en la ***Figura 4***.

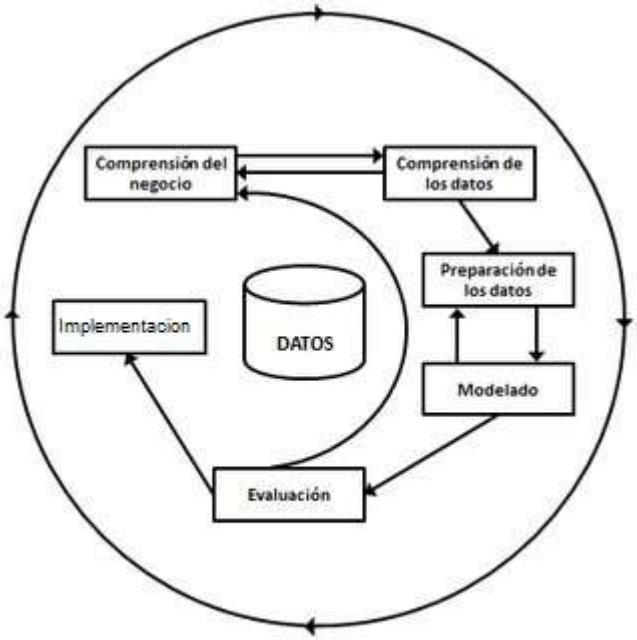

Figura 4. Fases de la metodología CRISP-DM

a. **La fase de comprensión del negocio:** La problemática que se centra en está fase es aquella comprensión de unos objetivos y requisitos que encamina al proyecto a un modelo empresarial, para convertir los conocimientos adquiridos en una denominación del problema de la minería de datos y así plantear un diseño para alcanzar objetivos

b. **La fase de comprensión de datos**: se propone identificar los problemas dentro de los datos e identificar temas de relevancia para formular hipótesis de información oculta.

Para ello partiendo de una colección inicial de datos y procesos con actividades busca familiarizarse con los datos.

c. **La fase de preparación de datos**: constituye un conjunto de actividades para elaborar un grupo de datos, está labor se ejecuta en múltiples oportunidades y sin tener un orden establecido. además, se incluye tareas de transformación de tablas, registros y atributos y limpieza de datos para las herramientas de modelado.

d. **La fase de modelado:** consiste en aplicar técnicas de modelado para obtener óptimos resultados que cumplan con unos requerimientos específicos sin perjuicio de que ello implique volver a la fase de preparación de datos.

e. **La fase de evaluación**: estudia aquellos modelos que constituyen el cumplimiento de aquellos criterios para logar el éxito, son interpretados por los resultados finales de las fases de explotación, teniendo en cuenta las acciones de los procesos de negocios.

3. MATERIALES Y MÉTODOS

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Como fuentes de información se utilizaron los datos que se encontraban disponibles, al momento de la investigación, en las bases de datos del ICFES de los resultados de los estudiantes que presentaron las pruebas Saber 9°. Los datos más actualizados eran del periodo comprendido entre los años 2014, 2015 y 2016. Para el descubrimiento de patrones asociados al desempeño académico en las pruebas Saber 9°, se construyó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta WEKA (Witten, Frank & Hall, 2011). Se escogió este modelo porque según la experiencia de algunos autores (Han & Kamber, 2001; Sattler & Dunemann, 2001; Timarán & Millán, 2006), para este tipo de proyectos, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender los resultados obtenidos. Por otra parte se realizaron pruebas con otros algoritmos como lo son Zero R y Part los cuales trae el programa WEKA, estos clasificadores fueron aplicados a cada uno de los repositorios finales al igual que los arboles de decisión j48 y se realizó una comparación entre estos resultados para posteriormente elegir los arboles de decisión puesto que muestran un mayor porcentaje de aciertos en algunos casos y en otros el mismo porcentaje (ver anexos). Además, la importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. Por otra parte, se escogió WEKA por ser una herramienta de minería de datos de software libre, distribuida bajo licencia GPL, que contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, tiene herramientas para la visualización de estos datos y provee una interfaz gráfica que unifica las herramientas para acceder fácilmente a sus funcionalidades (Calleja, 2010; García-Gutiérrez, 2016).

Para el descubrimiento de patrones, se aplicó la metodología CRISP-DM (Chapman et al., 2000; Villena-Román, 2016). En cuanto a las metodologías para desarrollar análisis de minería de datos y en un intento de normalización del proceso, de forma similar a como se hace en ingeniería para normalizar el proceso de desarrollo de software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Chapman et al., 2000; Villena-Román, 2016), y SEMMA (Sample, Explore, Modify, Model, and Assess) (Azevedo & Santos, 2008). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. Azevedo y Santos (2008), comparan

ambas implementaciones y llegan a la conclusión que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente.

En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA. La metodología CRISP-DM para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender está tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características (Azevedo & Santos, 2008). CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos (Hernández et al., 2005) y contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

En la fase de análisis del problema se identificó con exactitud la problemática que se solucionaría utilizando la minería de datos, esto permitió recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Villena-Román, 2016). En la fase de análisis de los datos se realizó la recolección inicial de datos, para establecer un primer contacto con el problema, familiarizarse con ellos, identificando su calidad y establecer las relaciones más evidentes que permitieron definir las primeras hipótesis. En la fase de preparación, se seleccionaron los datos a los cuales se les aplicaría una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Villena-Román, 2016).

En la fase de modelado se seleccionaron las técnicas de minería de datos más apropiadas para el proyecto. En la fase de evaluación se verificó si el modelo se ajusta a las necesidades establecidas en el proyecto. Se evaluaron los patrones encontrados con el fin de determinar su validez, remover los redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Finalmente, en la fase de implementación, se trató de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión del MEN, ICFES y de las instituciones gubernamentales y educativas que velan por la calidad de la educación en Colombia y difundir informes sobre el conocimiento extraído (Villena-Román, 2016). (Timaran-Pereira, Caicedo-Zambrano, & Hidalgo-Troya, 2019)

4. RESULTADOS

Teniendo en cuenta las fases de la metodología CRISP-DM los siguientes son los resultados de cada una de las etapas:

4.1 Comprensión del negocio

Se basa en el desempeño académico de los estudiantes de las instituciones educativas colombianas en las Pruebas Saber 9° del cual se pretende encontrar patrones asociados al buen o mal rendimiento de las mismas por medio de árboles de decisión, en aspectos socioeconómicos, académicos e institucionales. Para esto, en primera instancia se cuenta con la base de datos proporcionada por el ICFES de los años 2014, 2015 y 2016 en donde se evidencia los elementos mencionados anteriormente. Además, se profundizó en aspectos teóricos sobre rendimiento escolar, lineamientos de Pruebas Saber, factores asociados al desempeño académico, competencias genéricas y estructura de las Pruebas Saber.

4.2 Objetivo

Descubrir factores asociados al desempeño académico en las competencias que evalúan las Pruebas Saber 9° de los estudiantes pertenecientes a instituciones educativas colombianas, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES en los periodos 2014,2015 y 2016, a través de árboles de decisión con minería de datos, que permitan generar conocimiento encaminado a soportar las decisiones institucionales y gubernamentales para el mejoramiento de la calidad educativa.

4.3 Comprensión de los datos

En está fase, se identificó, recopiló y familiarizó con la información disponible en las bases de datos del ICFES, a cerca de los resultados en las pruebas Saber 9 aplicadas en los años 2014, 2015 y 2016 y sobre los datos socioeconómicos, académicos e institucionales de los estudiantes que presentaron está prueba a nivel nacional.

Por tener los archivos del ICFES diferentes estructuras, como son departamento, entidad territorial, establecimiento, municipio sede y valores plausibles, se procede inicialmente a la construcción de tres repositorios con ayuda del SGBD PostgreSQL denominados valores_plaudibles_2014, valores_plaudibles_2015, valores_plaudibles_2016 cuyo número de registros y atributos se muestran en la Tabla 3:

Tabla 3. Repositorio de valores plausibles.

Repositorio	Atributos	No. Registros
valores_plausibles_2014	47	577634
valores_plausibles_2015	47	536714
valores_plausibles_2016	47	571718
Total		1686066

Fuente: elaboración propia

Analizando los atributos de estos repositorios, se realizó la comparación de los mismos, entre cada año y así ver que atributos estaban presentes en un año y no estaban reportados en los demás años y viceversa como es el caso de competencias ciudadanas y ciencias Este análisis se puede mirar en la Tabla 4

Tabla 4. Valores plausibles por años

No.	Atributo	2014	2015	2016	Observaciones
1	Estu_consecutivo	x	x	x	está presente en todos los repositorios
2	Grupo	x	x	x	está presente en todos los repositorios
3	N	x	x	x	está presente en todos los repositorios
4	Estrato	x	x	x	está presente en todos los repositorios
5	Aplicación	x	x	x	está presente en todos los repositorios
6	Grado	x	x	x	está presente en todos los repositorios
7	Jornada	x	x	x	está presente en todos los repositorios
8	Establecimiento	x	x	x	está presente en todos los repositorios
9	EnteTerr	x	x	x	está presente en todos los repositorios
10	Departamento	x	x	x	está presente en todos los repositorios
11	Municipio	x	x	x	está presente en todos los repositorios
12	Género	x	x	x	está presente en todos los repositorios
13	Sector	x	x	x	está presente en todos los repositorios
14	Zona	x	x	x	está presente en todos los repositorios
15	ZonaStab	x	x	x	está presente en todos los repositorios

16	Calendario	x	x	x	está presente en todos los repositorios
17	Nivel	x	x	x	está presente en todos los repositorios
18	ModeloEdu	x	x	x	está presente en todos los repositorios
19	Disenso	x	x	x	está presente en todos los repositorios
20	Leng_copietas	x	x	x	está presente en todos los repositorios
21	Leng_weight	x	x	x	está presente en todos los repositorios
22	Leng_score1	x	x	x	está presente en todos los repositorios
23	Leng_score2	x	x	x	está presente en todos los repositorios
24	Leng_score3	x	x	x	está presente en todos los repositorios
25	Leng_score4	x	x	x	está presente en todos los repositorios
26	Leng_score5	x	x	x	está presente en todos los repositorios
27	Mate_copietas	x	x	x	está presente en todos los repositorios
28	Mate_weight	x	x	x	está presente en todos los repositorios
29	Mate_score1	x	x	x	está presente en todos los repositorios
30	Mate_score2	x	x	x	está presente en todos los repositorios
31	Mate_score3	x	x	x	está presente en todos los repositorios
32	Mate_score4	x	x	x	está presente en todos los repositorios
33	Mate_score5	x	x	x	está presente en todos los repositorios
34	Cien_copietas	x		x	está presente en 2014 y 2016
35	Cien_weight	x		x	está presente en 2014 y 2016
36	Cien_score1	x		x	está presente en 2014 y 2016
37	Cien_score2	x		x	está presente en 2014 y 2016
38	Cien_score3	x		x	está presente en 2014 y 2016
39	Cien_score4	x		x	está presente en 2014 y 2016
40	Cien_score5	x		x	está presente en 2014 y 2016
41	Comp_copietas		x		solo está presente en 2015
42	Comp_weight		x		solo está presente en 2015
43	Comp_score1		x		solo está presente en 2015
44	Comp_score2		x		solo está presente en 2015
45	Comp_score3		x		solo está presente en 2015
46	Comp_score4		x		solo está presente en 2015
47	Comp_score5		x		solo está presente en 2015

Fuente: elaboración propia

Posteriormente se crea un repositorio llamado a_vp_final el cual contiene la información de los tres repositorios anteriores el cual tiene 1686066 registros y 48 atributos. A continuación, se realiza una exploración de los datos a través del análisis de tendencias de desempeño académico en las 4 competencias genéricas estudiadas.

4.4 Tendencias de desempeño académico en competencias genéricas–pruebas saber 9.

Con el objetivo de explorar los datos se presenta en la tabla 4 un análisis de correlaciones entre las cuatro competencias genéricas de las pruebas Saber 9 (Matemáticas, Lenguaje, Ciencias y Competencias Ciudadanas) se estableció la tendencia está dística del desempeño académico en ellas con relación a aspectos socioeconómicos, académicos e institucionales. Para esto se utilizó la base de datos de las pruebas Saber 9 en los años 2014, 2015 y 2016 del ICFES y se seleccionaron los datos de valores plausibles; a través del coeficiente de correlación de Pearson, se establece cómo se asocian linealmente, entre sí, las cuatro competencias. Los resultados se presentan en la Tabla 5:

Tabla 5.Analisis de correlación entre las competencias genéricas.

Competencias Genéricas	Coeficiente de Correlación	Lenguaje	Matemáticas	Ciencias Naturales	Competencias Ciudadanas
Lenguaje	Correlación de Pearson	1	,990**	,986**	,666**
	P valor		0,000	0,000	0,000
	N	22663	10530	6140	4209
Matemáticas	Correlación de Pearson	,990**	1	,993**	,712**
	P valor	0,000		0,000	0,000
	N	10530	22601	6118	4170
Ciencias Naturales	Correlación de Pearson	,986**	,993**	1	.b
	P valor	0,000	0,000		
	N	6140	6118	12559	0
Competencias Ciudadanas	Correlación de Pearson	,666**	,712**	.b	1
	P valor	0,000	0,000		
	N	4209	4170	0	8578

Fuente. Elaboración propia

**. La correlación es significativa en el nivel 0,01 (2 colas).

b. No se puede calcular porque, como mínimo, una de las variables es constante.

De acuerdo a los resultados obtenidos en la tabla 5, se observa que Lenguaje con Matemáticas, Ciencias Naturales y Competencias Ciudadanas presentan correlaciones altas (r>0,5) por lo que se espera que los estudiantes que tienen buen desempeño en la prueba de Lenguaje también tengan buen desempeño en Matemáticas y Ciencias Naturales.

4.4.1 Desempeño en las cuatro competencias genéricas según variables socioeconómicas, académicas e institucionales.

Las tendencias de desempeño académico de los estudiantes y las cuatro competencias genéricas de las pruebas Saber 9, en relación con los aspectos socioeconómicos, académicos e institucionales se obtuvieron cruzando los puntajes de las competencias obtenidos en la prueba con las variables mencionadas anteriormente y se utilizaron las medias está dísticas: promedio, desviación estándar, intervalos de confianza y el tamaño del efecto de la diferencia estandarizada de las medias *d* de Cohen.

Para establecer si las diferencias en puntajes obtenidos en las cuatro competencias son está dísticamente significativas, se calcularon los intervalos de confianza de medias al 95%. Por otra parte, para determinar el tamaño de las diferencias de los promedios entre los diferentes grupos de variables analizadas se utilizó el estadístico d de Cohen con la siguiente escala propuesta por Cohen (1998): [0.0, 0.2] trivial o muy pequeña, [0.5, 0.8] moderada y [0.8, infinito] grande. Este estadístico se obtiene dividiendo la diferencia de medias (en valor absoluto) de los dos grupos por comparar entre la desviación estándar conjunta de estos. Para su cálculo se utiliza como referencia el grupo del más alto promedio en cada competencia. A continuación, se presenta el análisis de desempeño de los estudiantes en las pruebas Saber 9 en los años 2014, 2015 y 2016 según género, sector y tipo de establecimiento educativo, zona, nivel socioeconómico y jornada.

4.4.2 Género y Desempeño Académico en Competencias Genéricas

En la tabla 6 se muestra el desempeño en las competencias genéricas según el género. A partir de está tabla se concluye que los hombres presentan mejor desempeño que las mujeres

en tres de las competencias genéricas excepto en competencias ciudadanas en la cual sobresale el ítem donde no se especificó el género, sin embargo, según él estadístico d de Cohen estás diferencias son de magnitud trivial en las cuatro competencias.

Tabla 6. Desempeño académico en competencias genéricas según el género.

Competencia Genérica	Genero	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Lenguaje	Masculino	21814	6,12576	26,74052	0,18105	
	Femenino	22178	6,20570	26,99699	0,18128	0,00298
	No Especifica	12650	3,23221	15,54940	0,13825	0,12436
Matemáticas	Masculino	21679	6,02535	26,31492	0,17872	
	Femenino	22056	6,11467	26,55622	0,17881	0,00338
	No Especifica	12827	3,17830	15,43198	0,13626	0,12442
Ciencias Naturales	Masculino	12109	8,57988	33,60639	0,30540	
	Femenino	12294	8,90021	34,56868	0,31177	0,00940
	No Especifica	7254	4,54133	21,78061	0,25573	0,13583
Competencias Ciudadanas	Masculino	8290	1,83583	0,33848	0,00372	
	Femenino	8428	1,83422	0,33171	0,00361	0,00480
	No Especifica	5007	1,80993	0,28977	0,00410	0,08070

Fuente. **Elaboración propia**

4.4.3 Sector y desempeño académico en competencias genéricas

En la Tabla 7 se muestra el desempeño en las competencias genéricas según el sector del establecimiento educativo. A partir de está tabla se concluye que en las pruebas de competencias genéricas Saber 9 del 2014, 2015 y 2016 las instituciones no oficiales presentan

mejor desempeño que las instituciones oficiales, excepto en competencias ciudadanas en donde se observa una pequeña diferencia. Según él estadístico d de Cohen estás diferencias son de magnitud pequeña en las cuatro competencias.

Tabla 7.Desempeño académico en competencias según el sector del establecimiento.

Competencias Genéricas	Sector	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Lenguaje	Oficial	19829	1,8826	4,0354	0,0287	0,5553
	No Oficial	7698	15,7324	46,7134	0,5324	
Matemáticas	Oficial	19741	1,8558	2,9232	0,0208	0,5612
	No Oficial	7677	15,8221	46,7955	0,5341	
Ciencias Naturales	Oficial	10910	1,9499	4,5637	0,0437	0,6674
	No Oficial	4423	22,5747	57,0900	0,8584	
Competencias Genéricas	Oficial	7543	1,8441	0,3378	0,0039	0,1375
	No Oficial	2735	1,7969	0,3594	0,0069	

Fuente. Elaboración propia

4.4.4 Zona y desempeño académico en competencias genéricas.

A partir de la Tabla 8 se puede concluir que los establecimientos educativos ubicados en zona urbana presentan mejor desempeño que los establecimientos ubicados en zona rural, sin embargo, en competencias ciudadanas el promedio favorece a los establecimientos ubicados en zona rural. Según él estadístico d de Cohen estás diferencias son de magnitud trivial.

Tabla 8. Desempeño académico en competencias genéricas según la zona del establecimiento.

Competencias Genéricas	Zona	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Lenguaje	Urbano	611	35,9640	70,7749	2,8632	0,3963
	Rural	412	12,0485	40,1598	1,9785	
Matemáticas	Urbano	602	36,9983	72,0455	2,9364	0,3246
	Rural	414	16,5773	46,5557	2,2881	

Ciencias Naturales	Urbano	369	52,9404	85,0224	4,4261	0,5425
	Rural	267	14,3333	45,5252	2,7861	
Competencias Genéricas	Urbano	228	2,0833	0,5286	0,0350	0,1778
	Rural	166	2,1867	0,6477	0,0503	

Fuente. **Elaboración propia**

4.4.5 Nivel socioeconómico y desempeño académico en competencias genéricas

En la Tabla 9 se puede observar que en las competencias de lenguaje, matemáticas y ciencias el desempeño académico no tiene mayor diferencia entre los establecimientos pertenecientes a los diferentes estratos y en ciencias naturales sobresalen los establecimientos educativos pertenecientes a los estratos 1, 2 y 3. Según él estadístico d de Cohen estás diferencias son de magnitud pequeña en lenguaje, matemáticas y ciencias, y trivial en competencias ciudadanas.

Tabla 9.desempeño académico en competencias genéricas según el nivel socioeconómico del establecimiento.

Competencias Genéricas	Nivel Socioeconómico	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Lenguaje	Estrato 1	7043	2,0686	7,2162	0,0860	0,3442
	Estrato 2	10246	2,4147	10,9099	0,1078	0,3559
	Estrato 3	9352	2,6146	11,6413	0,1204	0,3401
	Estrato 4	8567	12,8009	41,5739	0,4492	
	Estrato 5	573	15,4799	47,5271	1,9855	0,0638
Matemáticas	Estrato 1	6987	2,0308	5,9011	0,0706	0,3482
	Estrato 2	10155	2,3917	10,3492	0,1027	0,3595
	Estrato 3	9358	2,5289	10,8336	0,1120	0,3471
	Estrato 4	8576	12,9266	41,8147	0,4515	
	Estrato 5	563	14,6371	47,5723	2,0049	0,0405
Ciencias Naturales	Estrato 1	3920	2,2428	8,4021	0,1342	0,4542
	Estrato 2	6264	2,6580	13,0119	0,1644	0,4874
	Estrato 3	5140	3,1529	15,1390	0,2112	0,4458
	Estrato 4	4369	19,9305	53,0435	0,8025	
	Estrato 5	239	30,3557	70,2800	4,5460	0,1928
Competencias Ciudadanas	Estrato 1	2578	1,8719	0,3948	0,0078	0,0668
	Estrato 2	3247	1,8479	0,3285	0,0058	0,1212
	Estrato 3	3383	1,8303	0,2988	0,0051	0,2355
	Estrato 4	3549	1,7892	0,3159	0,0053	0,0909
	Estrato 5	266	1,8355	0,4544	0,0279	

Fuente. Elaboración propia

4.4.6 Jornada y desempeño académico en competencias genéricas

En la Tabla 10 se evidencia que el desempeño académico en las competencias genéricas de lenguaje, matemáticas y ciencias es mayor en los establecimientos que pertenecen a jornada de la mañana. En competencias ciudadanas se evidencia que tienen mayor rendimiento los

establecimientos que pertenecen a la jornada mañana. Según el estadístico d de Cohen las diferencias son de magnitud trivial en las cuatro competencias.

Tabla 10.Desempeño académico en competencias genéricas según la jornada.

Competencias Genéricas	Jornada	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Leguaje	Completa	96764	1,6827	0,2192	0,0007	0,1063
	Mañana	616170	1,6992	0,2410	0,0003	0,0390
	Tarde	337411	1,7088	0,2524	0,0004	0,0000
	Única	5983	1,7078	0,1988	0,0026	0,0040
Matemáticas	Completa	70916	1,6947	0,2300	0,0009	0,0750
	Mañana	447972	1,7037	0,2482	0,0004	0,0399
	Tarde	241081	1,7138	0,2621	0,0005	0,0000
	Única	3804	1,7499	0,2319	0,0038	0,1378
Ciencias Naturales	Completa	62038	1,6561	0,1939	0,0008	0,1964
	Mañana	391439	1,6905	0,2312	0,0004	0,0530
	Tarde	216875	1,7032	0,2516	0,0005	0,0000
	Única	3896	1,7035	0,1710	0,0027	0,0011
Competencias Ciudadanas	Completa	33727	1,7907	0,2488	0,0014	0,1237
	Mañana	217651	1,7693	0,2528	0,0005	0,2067
	Tarde	116896	1,7752	0,2512	0,0007	0,1842
	Única	1877	1,8215	0,2582	0,0060	0,0000

Fuente. Elaboración propia

4.4.7 Calendario académico y desempeño académico en competencias genéricas.

Según la Tabla 11 el desempeño académico en las competencias genéricas de Lenguaje, Matemáticas, Ciencias Naturales y Competencias ciudadanas es mayor en el calendario A. Según el estadístico d de Cohen las diferencias entre las tres primeras competencias son de magnitud moderada y en Competencias Ciudadanas son de magnitud trivial.

Tabla 11. Desempeño académico en competencias genéricas según calendario académico.

Competencia Genérica	Calendario	N	Media	Desviación estándar	Media de error estándar	d de Cohen
Lenguaje	A	21878	1,8153	0,3616	0,0024	0,30370
	B	1631	68,5529	83,4174	2,0655	
Matemáticas	A	21791	1,8128	0,3602	0,0024	0,30376
	B	1634	68,3809	82,9659	2,0525	
Ciencias Naturales	A	11981	1,8461	0,3724	0,0034	0,39779
	B	967	98,1223	88,5537	2,8477	
Competencia Genérica	A	8480	1,8395	0,3447	0,0037	0,2773
	B	544	1,7441	0,3316	0,0142	

Fuente. Elaboración propia

Después de realizar el análisis de las tendencias de desempeño académico en competencias genéricas teniendo en cuenta el repositorio a_vp_final y obtener algunas conclusiones relevantes, se procede a organizar por grupos.

4.4.8 Descripción de diccionario de datos inicial.

El diccionario inicial contiene la descripción de cada una de las variables de las diferentes tablas como son: valores plausibles, instituciones, sedes, municipios, departamentos, entidades territoriales, entre otras. El repositorio inicial contenía tres bases de datos que se nombraron según el año de presentación de las pruebas saber 9 como: valores_plausibles_2014, valores_plausibles_2015 y valores_plausibles_2016; estás tres bases de datos se unificaron en un solo repositorio llamado a_vp_final_1 el este cuenta con 1.683.063 registros y 47 atributos. En la Tabla 12 se muestra el diccionario de datos de los valores plausibles (en adelante estudiantes), que incluye el nombre del campo, la descripción de las variables y el tipo de campo.

Tabla 12. Diccionario de datos de valores plausibles (Estudiantes).

Nombre de Campo	Descripción	Tipo de Campo
Estu_consecutivo	Código de identificación de hoja de respuestas	VAAAAGXXXXXX donde V = constante AAAA = Año, G = grado XXXXXXX = consecutivo
Grupo	Grupo/salón al que pertenecen los estudiantes	99 = censal ¬99 = control
N	Matricula del grupo al que pertenecen	1, 2, …,
Estrato	Estrato a que pertenece el establecimiento dentro del arco muestral	Alfanumérico(9)
Aplicación	No definido	X
Grado	Grado del estudiante	3, 5, 9.
Jornada	(Codsitio) - código de la sede - jornada al que pertenece el estudiante ()	Numérico (6) ir a la tabla sedes en la columna ID
Establecimiento	Código DANE del establecimiento educativo al que pertenece el estudiante	Llave a tabla INSTITUCION
EnteTerr	Código DANE de la entidad a la que pertenece establecimiento educativo	Llave a tabla ENTIDADTERRITORIAL
Departamento	Código DANE del departamento al que pertenece establecimiento educativo	Llave a tabla DEPARTAMENTO
Municipio	Código DANE del municipio al que pertenece el establecimiento educativo	Llave a tabla MUNICIPIO
Género	Sexo del estudiante	1=Masculino 2=Femenino
		3= No especifica

Sector	Sector del establecimiento	1 = Oficial 2 = No oficial
Zona	Zona donde se ubica la mayoría de la población atendida	1 = Urbana 2 = Rural
ZonaStab	Zona del establecimiento	1 = Urbana 2 = Rural
Calendario	Calendario del establecimiento	A B
Nivel	Nivel socioeconómico del establecimiento	1 2 3 4 5
ModeloEdu	No definido	X
Disenso	Marca de discapacidad cognitiva	0 = sin discapacidad 1 = con discapacidad
Leng_copietas	Indicador de copia de lenguaje	0 = no copia 1 = con copia
Leng_weight	Peso muestral de lenguaje	Numérico(6)
Leng_score1	Score1 - Valor plausible 1 de lenguaje	Numérico(6)
Leng_score2	Score2 - Valor plausible 2 de lenguaje	Numérico(6)
Leng_score3	Score3 - Valor plausible 3 de lenguaje	Numérico(6)
Leng_score4	Score4 - Valor plausible 4 de lenguaje	Numérico(6)
Leng_score5	Score5 - Valor plausible 5 de lenguaje	Numérico(6)
Mate_copietas	Indicador de copia de matemáticas	0 = no copia 1 = con copia
Mate_weight	Peso muestral de matemáticas	Numérico(6)

Mate_score1	Score1 - Valor plausible 1 de matemáticas	Numérico(6)
Mate_score2	Score2 - Valor plausible 2 de matemáticas	Numérico(6)
Mate_score3	Score3 - Valor plausible 3 de matemáticas	Numérico(6)
Mate_score4	Score4 - Valor plausible 4 de matemáticas	Numérico(6)
Mate_score5	Score5 - Valor plausible 5 de matemáticas	Numérico(6)
Cien_copietas	Indicador de copia de ciencias	0 = no copia 1 = con copia
Cien_weight	Peso muestral de ciencias	Numérico(6)
Cien_score1	Score1 - Valor plausible 1 de ciencias	Numérico(6)
Cien_score2	Score2 - Valor plausible 2 de ciencias	Numérico(6)
Cien_score3	Score3 - Valor plausible 3 de ciencias	Numérico(6)
Cien_score4	Score4 - Valor plausible 4 de ciencias	Numérico(6)
Cien_score5	Score5 - Valor plausible 5 de ciencias	Numérico(6)
Comp_copietas	Indicador de copia de competencias ciudadanas	0 = no copia 1 = con copia
Comp_weight	Peso muestral de competencias	Numérico(6)
Comp_score1	Score1 - Valor plausible 1 de competencias ciudadanas	Numérico(6)
Comp_score2	Score2 - Valor plausible 2 de competencias ciudadanas	Numérico(6)
Comp_score3	Score3 - Valor plausible 3 de competencias ciudadanas	Numérico(6)
Comp_score4	Score4 - Valor plausible 4 de competencias ciudadanas	Numérico(6)

Comp_score5	Score5 - Valor plausible 5 de competencias ciudadanas	Numérico(6)
Año	Año que se presentó la prueba	Numérico (4)

Fuente. **Elaboración propia**

La base de datos cuenta con tablas adicionales por establecimiento educativo como son instituciones completas e instituciones simplificadas que se utilizaran para limpiar de mejor manera los repositorios.

En la .

Tabla 13 se indica los reportes completos que son usados para indicar las está dísticas de cada establecimiento donde, por cada grado área participaron seis o más estudiantes en la prueba.

Tabla 13. Resultado de instituciones completo.

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que participan en la evaluación
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia, 1=Indicios de copia individual 2=Indicios de copia masiva en alguna sede – jornada
Peso	Numérica	Peso o ponderación a usar para el cálculo de agregados con la información a nivel de sede jornada
Promedio	Numérica	promedio de los puntajes de los estudiantes dentro del establecimiento educativo
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia, 1=Indicios de copia individual, 2=Indicios de copia masiva en alguna sede – jornada

Peso	Numérica	peso o ponderación a usar para el cálculo de agregados con la información nivel de sede – jornada
Error Estándar	Numérica	El error estándar del promedio de los puntajes de los estudiantes dentro del establecimiento Educativo
Desviación	Numérica	Desviación estándar del puntaje de los estudiantes dentro del establecimiento Educativo
Insuficiente	Numérica	Porcentaje de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Porcentaje de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Porcentaje de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Porcentaje de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

En la Tabla 14 se encuentra la descripción de variables de los reportes simplificados que pertenecen a los establecimientos educativos en los que participaron menos de seis estudiantes por grado área.

Tabla 14.Resultados instituciones simplificado.

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que participa en la evaluación
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron evaluados en el área
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia 1=Indicios de copia individual 2=Indicios de copia masiva en alguna sede-jornada
Peso	Numérica	Peso o ponderación a usar para el cálculo de agregados con la información a nivel sede – jornada

Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

La

Tabla ***15*** muestra los resultados simplificados a nivel de agregación de la sede jornada e indican el número de estudiantes por cada nivel de desempeño.

Tabla 15. Resultado sede - jornada.

Nombre de Campo	Tipo de Campo	Descripción
Id_Sede	Texto	Código saber de la sede jornada
Codigo_Dane_Sede	Texto	Código DANE de la sede del establecimiento Educativo
Jornada	Texto	Jornada en la que desempeña actividades y para la que se reporta resultados. M=Mañana, T=Tarde, C=Completa
Evaluados	Numérica	Total de estudiantes del establecimiento educativo que participan en la evaluación
Participantes	Numérica	Total de estudiantes del establecimiento educativo que fueron evaluados en el área
Copia	Texto	Indicio de copia en el área: 0=No se presentan indicios de copia 1=Indicios de copia individual, 2=Indicios de copia masiva en alguna sede-jornada
Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo

Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado

Fuente. Tomado de las bases de datos ICFES.

La ño.

Tabla 16 muestra los resultados simplificados a nivel de agregación de municipio e indican el número de estudiantes por cada nivel de desempeño.

Tabla 16. Resultados municipio.

Nombre de Campo	Tipo de Campo	Descripción
Muni_Id	Texto	Código saber del municipio
Municipio	Texto	Nombre del municipio según divipo
Departamento	Texto	Nombre del departamento
Puntaje_Promedio	Numérica	Promedio de los puntajes de los estudiantes dentro del establecimiento Educativo
Error estándar Promedio	Numérica	El error estándar del promedio de los puntajes de los estudiantes dentro del establecimiento Educativo
Desviación	Numérica	Desviación estándar del puntaje de los estudiantes dentro del establecimiento Educativo
Insuficiente	Numérica	Número de estudiantes en el nivel de desempeño insuficiente
Mínimo	Numérica	Número de estudiantes en el nivel de desempeño mínimo
Satisfactorio	Numérica	Número de estudiantes en el nivel de desempeño satisfactorio
Avanzado	Numérica	Número de estudiantes en el nivel de desempeño avanzado
N	Numérica	Número de participantes

Fuente. Tomado de las bases de datos ICFES.

A continuación, se muestran tablas adicionales de identificación (

Tabla ***17***hasta Tabla 20) que contienen información sobre entidad territorial, municipio, departamentos, establecimiento educativo y sedes.

Tabla 17. Identificación del campo entidades.

Nombre de Campo	Tipo de Campo	Descripción
Id_Ente	Texto	Identificador del ente territorial
Nombre	Texto	Nombre del ente territorial
Munexclu	Texto	Nombre de los municipios certificados excluidos dentro del ente Territorial
Tipo	Texto	Tipo de ente territorial 1=ETC,2=DPTO,3=MPIO

Fuente. Tomado de las bases de datos ICFES.

Tabla 18. Identificación del campo municipios.

Nombre de Campo	Tipo de Campo	Descripción
Id Municipio	Texto	Código dane del municipio según divipo
Nombre	Texto	Nombre del municipio según divipo

Fuente. Tomado de las bases de datos ICFES

Tabla 19. Identificación del campo establecimientos.

Nombre de Campo	Tipo de Campo	Descripción
Cod_Dane	Texto	Código DANE del establecimiento educativo
Id_Municipio	Texto	Código dane del municipio al que pertenece el establecimiento Educativo

Id_Ente	Texto	Id de la entidad territorial
Nombre	Texto	Nombre del establecimiento educativo reportado en el DUE
Zona	Texto	Zona donde la mayoría de la población atendida por el establecimiento educativo se ubica. 1=Urbano, 2=rural
Sector	Texto	Naturaleza administrativa del establecimiento educativo. 1= Oficial,2=No oficial
Tipo_Está b	Texto	Tipo de establecimiento. 1=Oficial urbano, 2=Oficial rural, 3=No Oficial
Calendario	Texto	Calendario del establecimiento
Nivel _Socio	Texto	NSE asignado de acuerdo a la clasificación realizada con puntajes Promedios

Fuente. Tomado de las bases de datos ICFES

Tabla 20. Identificación del campo sedes.

Nombre de Campo	Tipo de Campo	Descripción
Id	Texto	Código saber de la sede jornada
Codigo_Dane_Estab	Texto	Código DANE del establecimiento educativo
Codigo_Dane_Sede	Texto	Código DANE de la sede del establecimiento educativo
Jornada	Texto	Jornada en la que desempeña actividades y para la que se reporta resultados. M=Mañana, T=Tarde, C=Completa
Nombre	Texto	Nombre de la sede jornada según DUE

Fuente. Tomado de las bases de datos ICFES.

Las tablas que se muestran a continuación (Tabla 21 hasta Tabla 30) fueron adicionadas por conveniencia ya que en ellas se encontraban datos implícitos como valores de 0 y 1, iniciales de palabras, entre otras y se renombró con caracteres.

Tabla 21. Descripción inicio de copia.

ID	Descripción
0	No se presentan indicios de copia
1	Indicios de copia individual
2	Indicios de copia masiva en alguna sede-jornada

Fuente. Elaboración propia

Tabla 22. Descripción de jornada.

ID	Descripción
M	Mañana
T	Tarde
C	Completa
U	Única

***Fuente.* Elaboración propia**

Tabla 23. Descripción de tipo de entidad.

ID	Descripción
0	Entidad Territorial
1	Departamento
2	Municipio
4	Municipio si
5	No certificadas

***Fuente.* Elaboración propia**

Tabla 24. Descripción de la zona.

ID	Descripción
1	Urbano
2	Rural

Fuente. Elaboración propia

Tabla 25. Descripción de capacidad.

ID	Descripción
0	Sin Discapacidad
1	Con Discapacidad

***Fuente.* Elaboración propia**

Tabla 26. Descripción del sector.

ID	Descripción
1	Oficial
2	No Oficial

***Fuente.* Elaboración propia**

Tabla 27. Descripción del tipo de establecimiento.

ID	Descripción
1	Oficial Urbano
2	Oficial Rural
3	No Oficial

Fuente. Elaboración propia

Tabla 28. Descripción de género.

ID	Descripción
1	Masculino
2	Femenino
3	No Especifica

Fuente. Elaboración propia

Tabla 29. Descripción de copietas.

ID	Descripción
0	No Copia
1	Con Copia

Fuente. Elaboración propia

4.5 Preparación de los datos

Teniendo en cuenta la base de datos proporcionada por el ICFES de las pruebas saber 9 la cual contiene varios repositorios como son departamento, entidad territorial, establecimiento, municipio, sede y valores plausibles para los años 2014, 2015 y 2016 se procedió a la depuración de los mismos teniendo en cuenta la calidad de los datos y las técnicas de Minería que se va aplicar. Además, se tiene en cuenta los repositorios de valores plausibles que son los resultados de las pruebas individuales en cada año e incluyen todos los datos que contienen los diferentes repositorios por separado.

A continuación, se describe los procesos realizados en las fases de limpieza y transformación de datos.

4.5.1 Limpieza

Uno de los requerimientos para aplicar las técnicas de minería de datos es que el repositorio esté limpio, es decir, que no exista presencia de datos faltantes o perdidos (missing values) o valores que no se ajusten al comportamiento general de los datos (outliers), para esto se ve la necesidad de mejorar la calidad de los repositorios iniciales los cuales se caracterizan por tener un esquema asociado, es decir los datos siguen una estructura y son por lo tanto estructurados. Lo cual permite una consulta donde se puede combinar en una sola tabla la información de varias tablas que se requieran para cada tarea concreta, este proceso se realizó con el programa Postgrest SQL,

Así se crearon los repositorios a_vp_final_1, del repositorio a_vp_final_1 se depuraron algunos datos nulos que se encontraban en las bases de datos como se especifican en la Tabla 30.

Tabla 30. Atributos con alto porcentaje de valores nulos.

Año	Atributo	Participantes por Materia	Nulos	Porcentaje %
2014	leng_copietas leng_weight, leng_score1 leng_score2, leng_score3, leng_score4 y leng_score5	379401	198232	34.32
	mate_copietas mate_weight, mate_score1 mate_score2, mate_score3, mate_score4 y mate_score5	377685	199948	34.62
	cien_copietas, cien_weight, cien_score1 cien_score2, cien_score3, cien_score4 y cien_score5	363687	213946	37.04
	comp_copietas com_weight, comp_score1 comp_score2, comp_score3 comp_score4 y comp_score5	0	577633	100.00
2015	leng_copietas leng_weight, leng_score1 leng_score2, leng_score3, leng_score4 y leng_score5	356191	180522	33.63

	mate_copietas mate_weight, mate_score1 mate_score2, mate_score3, mate_score4 y mate_score5	353970	182743	34.05
	cien_copietas, cien_weight, cien_score1 cien_score2, cien_score3, cien_score4 y cien_score5	0	536713	100.00
	comp_copietas com_weight, comp_score1 comp_score2, comp_score3 comp_score4 y comp_score5	356908	179805	33.50
2016	leng_copietas leng_weight, leng_score1 leng_score2, leng_score3, leng_score4 y leng_score5	380282	191435	33.48
	mate_copietas mate_weight, mate_score1 mate_score2, mate_score3, mate_score4 y mate_score5	378791	192926	33.75
	cien_copietas, cien_weight, cien_score1 cien_score2, cien_score3, cien_score4 y cien_score5	375665	196052	34.29
	comp_copietas com_weight, comp_score1 comp_score2, comp_score3 comp_score4 y comp_score5	0	571717	100.00

Fuente. Elaboración propia

Los porcentajes del 100% de valores nulos se deben a que no se aplicó la competencia de competencias ciudadanas en el año 2015 y la competencia de ciudadanía en los años 2014 y 2016. Además, los porcentajes que se aproximan al 35% que se indican en las variables,

corresponden a aquellos estudiantes que sólo aplican dos de las cuatro competencias genéricas mencionadas anteriormente.

Después de esto se procede a la eliminación de atributos que son relevantes para la investigación como se puede ver en la Tabla 31 eliminar atributos. [1]

Tabla 31.Atributos eliminados.

Atributo	descripción	cantidad de eliminados
estu_consecutivo	Contiene valores constantes sirve para identificación del estudiante.	1686063
grupo		1686063
n	Muestra el número de participantes que presentaron la prueba.	1686063
aplicación	Contenía la letra "X" que significaba que si presentaron la prueba.	1686063
grado	El valor era constante puesto que todos los estudiantes son de grado 9.	1686063
modelo educativo	Las instituciones no tienen definido un modelo educativo.	1686063
disenso	Contiene datos de estudiantes con discapacidad, se puede ver que el número de estudiantes con discapacidad es muy pequeño.	1686063
mate_score	Se refiere a la competencia de matemática, está se divide en 5 variables (mate_score1, mate_score2, mate_score3, mate_score4, mate_score5) y los valores no se encuentran en el rango (-3 a 3) según el ICFES (2011), también presentan celdas vacías.	577617

[1] "Para los estudiantes que presentaron la prueba SABER 3°, 5°. y 9°, se obtuvo el resultado de su desempeño a partir de cinco valores, denominados valores plausibles (PV1 a PV5) o como están en el FTP (score1 a score5) que generalmente vienen en una escala de -3 a 3. Estos valores son útiles, pues tienen en cuenta la aleatoriedad producida por el hecho de que los estudiantes responden a un número pequeño de preguntas, lo cual permite obtener mejores estimaciones de las estadísticas de interés relacionadas con el desempeño en las pruebas a nivel agregado".(ICFES,2011)

leng_score	Se refiere a la competencia de matemáticas, está se divide en 5 variables (leng_score1, leng_score2, leng_score3, leng_score4, leng_score5) y los valores no se encuentran en el rango (-3 a 3) según el ICFES (2011), también presentan celdas vacías.	570189
comp_score	Se refiere a la competición de competencias ciudadanas, está se divide en 5 variables (comp_score1, comp_score2, comp_score3, comp_score4, comp_score5) y los valores no se encuentran en el rango (-3 a 3) según el ICFES (2011), también presentan celdas vacías.	1329155
cien_score	Se refiere a la competencia de ciencias naturales, está se divide en 5 variables (cien_score1, cien_score2, cien_score3, cien_score4, cien_score5) y los valores no se encuentran en el rango (-3 a 3) según el ICFES (2011), también presentan celdas vacías.	946711

Fuente. Elaboración propia

Posteriormente se procede a crear una tabla que contenga las competencias genéricas que se hayan aplicado por año en las pruebas saber 9 como se puede apreciar en la Tabla 32.

Tabla 32. Posibles combinaciones entre las competencias genéricas.

No.	Lenguaje	Matemáticas	Ciencias	Competencias	No. Datos
1	1	0	0	1	177414
2	1	1	0	1	0
3	0	1	1	1	0

4	0	0	0	1	2892
5	1	1	1	1	0
6	1	0	0	0	546315
7	1	1	1	0	0
8	1	0	1	0	368902
9	0	0	1	1	0
10	0	1	0	1	176602
11	0	1	1	0	367284
12	0	0	1	0	3166
13	1	1	0	0	546315
14	0	1	0	1	69497
15	1	1	1	1	0
16	0	0	0	0	0

Fuente. Elaboración propia

En la tabla hay 16 filas las cuales corresponden a las posibles combinaciones de las competencias genéricas, se tienen en cuenta aquellas que se pueden dar , como son: las filas 1 muestra la aplicación de las competencias de lenguaje y competencias ciudadanas con 177414 datos, la fila 8 muestra la aplicación de las competencias de lenguaje y ciencias con 368902 datos , la fila 9 se descarta aunque hay dos competencias genéricas como ciencias y competencias ciudadanas que no tienen ningún dato, la fila 11 muestra la aplicación de las competencias matemáticas y ciencias con 367284 datos, la fila 13 muestra la aplicación de las competencias de lenguaje y matemáticas con 546315 datos y la fila 15 muestra la aplicación de las competencias de matemáticas y competencias ciudadanas con 68497 datos.

4.5.2 Transformación

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso, se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos (Fayyad et al., 1996).

Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontalmente o verticalmente. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos (ejemplo, edad por un rango de edades).

La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (ejemplo, edad y fecha de nacimiento), se utilizan técnicas de reducción tales como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras (Timaran, Calderon, & Jimenez, 2013a).

Teniendo en cuenta los criterios anteriores en el repositorio a_vp_final_1 se eliminaron algunos atributos porque presentaban inconsistencias o eran relevantes para la investigación por ejemplo grado que es una constante puesto que todos los estudiantes son de grado 9 entre otros, como se explicó en la tabla de eliminación de atributos.

Para realizar la transformación de los datos se procede a crear el repositorio zona geográfica, este se crea teniendo en cuenta órganos colegiados de administración y decisión OCAD los cuales agrupan departamentos completos con sus respectivos municipios en diferentes regiones, teniendo en cuenta está organización es posible predecir cuál será el rendimiento académico en competencias genéricas en las pruebas saber 9.

Las OCAD están distribuidas en 6 zonas en todo el país, para está investigación se tiene en cuenta la ciudad de Bogotá como otra zona, porque en está ciudad se concentra gran cantidad de instituciones, así se tiene que hay 7 zonas geográficas como se muestra en la Tabla 33.

Tabla 33. Zona geográfica.

Zona Geográfica	Departamento
Bogotá	Bogotá Distrito Capital
Caribe	Atlántico, Bolívar, Cesar, Córdoba, La Guajira, San Andrés y Providencia, Magdalena y Sucre
Centro Sur	Amazonas, Caquetá, Huila, Putumayo y Tolima
Centro Oriente	Boyacá, Cundinamarca, Norte de Santander y Santander
Eje Cafetero	Antioquia, Caldas, Risaralda y Quindío
Llano	Arauca, Casanare, Guainía, Guaviare, Meta, Vaupés y Vichada
Pacifico	Cauca, Choco, Valle del Cauca y Nariño

Fuente. Elaboración propia

Una vez clasificadas las zonas geográficas se procede a clasificar el número de instituciones por zona geográfica y a organizarlas teniendo en cuenta: si el número de estudiantes es mayor a 300.000 entonces se clasifica como alto, si el número de estudiantes está entre 200.000 y 300.000 se clasifica como medio y se clasifica como bajo si el número de estudiantes es menor a 200.000 en la Tabla 34 se pueden ver los resultados.

Tabla 34. Clasificación de estudiantes por zona.

Zona	Número de estudiantes	Clasificación
Llano	67081	Bajo
Centro sur	121843	Bajo
Bogotá	274635	Medio
Caribe	396196	Alto
Centro Oriente	286393	Medio
Eje Cafetero	280603	Medio
Pacifico	257962	Medio

Fuente. Elaboración propia

También se tiene en cuenta el número de instituciones educativas que tiene cada zona geográfica y se las clasifica teniendo en cuenta si el número de instituciones es mayor a 4000 se clasifica como alto, si el número de instituciones está entre 2000 y 4000 se clasifica en medio y si clasifican en bajo es porque el número de instituciones es menor a 2000 en la

Tabla ***35*** se observan los resultados

Tabla 35. Clasificación de instituciones por zona.

Zona	Numero de instituciones	Clasificación
Llano	1371	Bajo
Centro sur	2663	Medio
Bogotá	3291	Medio
Caribe	6695	Alto
Centro Oriente	5620	Alto
Eje Cafetero	5157	Alto
Pacifico	5197	Alto

Fuente. Elaboración propia

Para transformar el repositorio a_vp_final1 se procedió a dividir este en las diferentes competencias como se mostró en la Tabla 32 obteniendo los siguientes repositorios ilustrados en la Tabla 36.

Tabla 36.Repositorios finales.

Repositorio	Registros	Atributos
a_mate_comp_final	69497	16

a_leng_comp_final	70167	16
a_mate_cien_final	156740	16
a_leng_cien_final	157640	16
a_leng_mate_final	227207	16
a_cien_comp_final	0	16

Fuente. Elaboración propia

Cada uno de estos repositorios tenían los atributos codificados en forma numérica para poder transformarlos en atributos alfanuméricos se procedió a crear tablas auxiliares donde se unifican los diferentes años 2014, 2015 y 2016 en un solo, también se les realizo un proceso de limpieza para que no hallan datos repetidos, con ayuda del programa post gres se llamó a las tablas auxiliares y así se pueda realizar los cambios. En la tabla se puede ver los nuevos repositorios.

Tabla 37. Repositorios Auxiliares.

Repositorio	Registros	Atributos
deptos_final	99	3
cod_municipios_total	1112	2
jornada_sede_total	95845	5
instituciones_jornada	62416	2

Fuente. Elaboración propia

Una vez realizado el proceso de transformación se obtuvo los repositorios finales apreciables en la siguiente **¡Error! La autoreferencia al marcador no es válida.**.

Tabla 38. Repositorios finales.

Atributo	Nuevo Atributo	Descripción	Acción Realizada	Valores
Jornada	Jornada	código de la sede-jornada a la que pertenecen los estudiantes	se cambia valor numérico por carácter	C: completa M: mañana T: tarde U: única
cod_dane	Nombre	código dane de la institución educativa	se cambia el código dane por el nombre de la institución educativa	Nombre
id_entidad	Entidad	código de la entidad territorial	se cambia el código de la entidad territorial por el nombre	Entidad
id_departamento	Departamento	código del departamento	se cambia el código del departamento por el nombre del departamento	nombre departamentos
muni_id	muni_nombre	código del municipio	se cambia el código del municipio por el nombre del municipio	nombre municipio
calendario	Calendario	calendario al que pertenece la institución educativa	no se realiza cambios	A, B
nivel	nivel	nivel socioeconómico del establecimiento	no se realiza cambios	1,2,3,4,5
año	Año	año en que se presentaron las pruebas saber 9	no se realiza cambios	2014 2015 2016
sexo	sexo_cuali	genero del estudiante	se remplaza por el genero	M: Masculino F: Femenino N: No especifica

sector	sectot_cuali	sector del establecimiento educativo	se cambia valor numérico por carácter	Oficial No Oficial
zona	zona_cuali	zona de la sede jornada	se cambia valor numérico por carácter	urbana rural
zonastab	zonastab_cuali	zona del establecimiento educativo	se cambia el valor numérico por carácter	urbana rural
zona_geo	zona_geo	zona geográfica a la cual pertenece el establecimiento educativo	se cambia el valor numérico por carácter	Caribe Centro sur Centro oriente Eje cafetero Llano Pacifico Bogotá
cien_weight	cien_weight	puntaje obtenido en la prueba saber 9,estos puntajes está n en valores numéricos	se cambian comas por puntos	mínimo máximo
leng_weight	leng_weight	puntaje obtenido en la prueba saber 9,estos puntajes está n en valores numéricos	se cambian comas por puntos	mínimo máximo
cien_weight_ normal	cien_weight_no rmal	puntaje obtenido en la prueba saber 9,estos puntajes está n en valores numéricos	se normalizan los puntajes obtenidos	mínimo máximo
leng_weight_ normal	leng_weight_no rmal	puntaje obtenido en la prueba saber 9,estos puntajes está n en valores numéricos	se normalizan los puntajes obtenidos	mínimo máximo
rendi_leng_n ormal	rendi_leng_nor mal	valores obtenidos en la competencia	se cambia el valor numérico por carácter	Insuficiente Medio bajo Bajo

				Medio Alto
rendi_leng_normal	rendi_leng_normal	valores obtenidos en la competencia	se cambia el valor numérico por carácter	Insuficiente Medio bajo Bajo Medio Alto
rendi_leng_cien	rendi_leng_cien_normal	valores obtenidos en la competencia	se cambia el valor numérico por carácter	Insuficiente Medio bajo Bajo Medio Alto

Fuente. Elaboración propia

Según los parámetros que maneja el ICFES la prueba evalúa dos de las cuatro competencias genéricas a cada estudiante, por lo tanto, como se mencionó antes, se construyeron 5 repositorios los cuales contienen el mismo número de atributos de la tabla anterior la única diferencia son las competencias que se evalúan en cada una.

En la siguiente Tabla 39 se muestran las características del conjunto de datos que constituyen los repositorios finales para aplicarles las técnicas de minería de datos.

Tabla 39. Características de los repositorios finales.

Repositorio	**No. Registros**	**No. Atributos**	**Descripción**
a_leng_comp_final_final	70167	20	conjuntos de datos de los estudiantes que presentaron la prueba de lenguaje y competencias ciudadanas
a_mate_comp_final_final	69497	20	conjuntos de datos de los estudiantes que presentaron

			la prueba de matemáticas y competencias ciudadanas
a_mate_cien_final_final	156740	20	conjuntos de datos de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales
a_leng_cien_final_final	157640	20	conjuntos de datos de los estudiantes que presentaron la prueba de lenguaje y ciencias naturales
a_leng_mate_final_final	227207	20	conjuntos de datos de los estudiantes que presentaron la prueba de lenguaje y matemáticas

Fuente. **Elaboración propia**

Finalmente se realizó la normalización de cada competencia individualmente para ello se procedió a dividir entre 100 los puntajes obtenidos puesto que en algunos casos no estaban dados como porcentaje sino como puntaje; esto no permitía realizar la clasificación en la escala de insuficiente, medio bajo, bajo, medio y alto una vez finalizado este paso se toma el puntaje porcentual mínimo y el puntaje porcentual máximo de cada una de las competencias para encontrar su rango para posteriormente dividirlo en cinco partes iguales y así poder realizar la clasificación antes mencionada una vez realizado este proceso se lo repite nuevamente pero está vez teniendo en cuenta las posibles combinaciones (binas) de las competencias que se están evaluando para finalmente dar la clasificación de está s.

A continuación se muestra el proceso en las competencias de lenguaje y matemáticas; este proceso fue realizado con todas las posibles combinaciones de competencias genéricas. (Ver Tabla 32).

Tabla 40. Límites para leng_weight_normal

Maxi	mini	rango	rango 1	rango 2	rango 3	rango 4
5	1	0,8	1,8	2,6	3,4	4,2

Fuente. **Elaboración propia**

Tabla 41. Clasificación leng_weight_normal

rendi_leng_normal	count
MEDIO BAJO	45178
INSUFICIENTE	180123
MEDIO	61
BAJO	1843
ALTO	1

Fuente. Elaboración propia

Tabla 42. Límites para mate_weight_normal

Maxi	mini	rango	rango 1	rango 2	rango 3	rango 4
6,0	1,0	1,0	2,0	3,0	4,0	5,0

Fuente. Elaboración propia

Tabla 43. Clasificación mate_weight_normal

rendi_mate_normal	count
MEDIO BAJO	21127
INSUFICIENTE	205598
BAJO	477
ALTO	4

Fuente. Elaboración propia

Tabla 44.Limites para rendi_leng_mate_weight_normal

Maxi	mini	rango	rango 1	rango 2	rango 3	rango 4
4,0	1,0	0,6	1,6	2,2	2,8	3,4

Fuente. **Elaboración propia**

Tabla 45. Clasificación rendi_leng_mate_normal

rendi_leng_mate_normal	count
ALTO	9
MEDIO	945
MEDIO BAJO	111731
BAJO	5647
INSUFICIENTE	108874

Fuente. Elaboración propia

4.6 Modelado

En está fase se seleccionó la tarea de clasificación con árboles de decisión como la técnica de minería de datos más adecuada para solucionar el problema objeto de la investigación. La clasificación es una técnica predictiva que se aplica a problemas donde hay que predecir nuevos datos para uno o más ejemplos que van acompañados de una salida denominada clase (Hernández et al., 2005). Con está clasificación se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes que presenten las pruebas saber 9 en todo el país teniendo en cuenta, factores como la jornada del establecimiento educativo, el calendario, la zona geográfica, el sexo, entre otros asociados a un probable bueno o mal desempeño académico en las cuatro competencias genéricas, evaluadas en las pruebas Saber 9.

Para el descubrimiento de patrones de desempeño académico en las competencias genéricas de los estudiantes de todo el país que presentaron las pruebas Saber 9 entre los años 2014 al 2016 se utilizó el software WEKA , una herramienta de minería de datos.

En la **Figura 5** se muestra el proceso de obtener el modelo de clasificación con está herramienta.

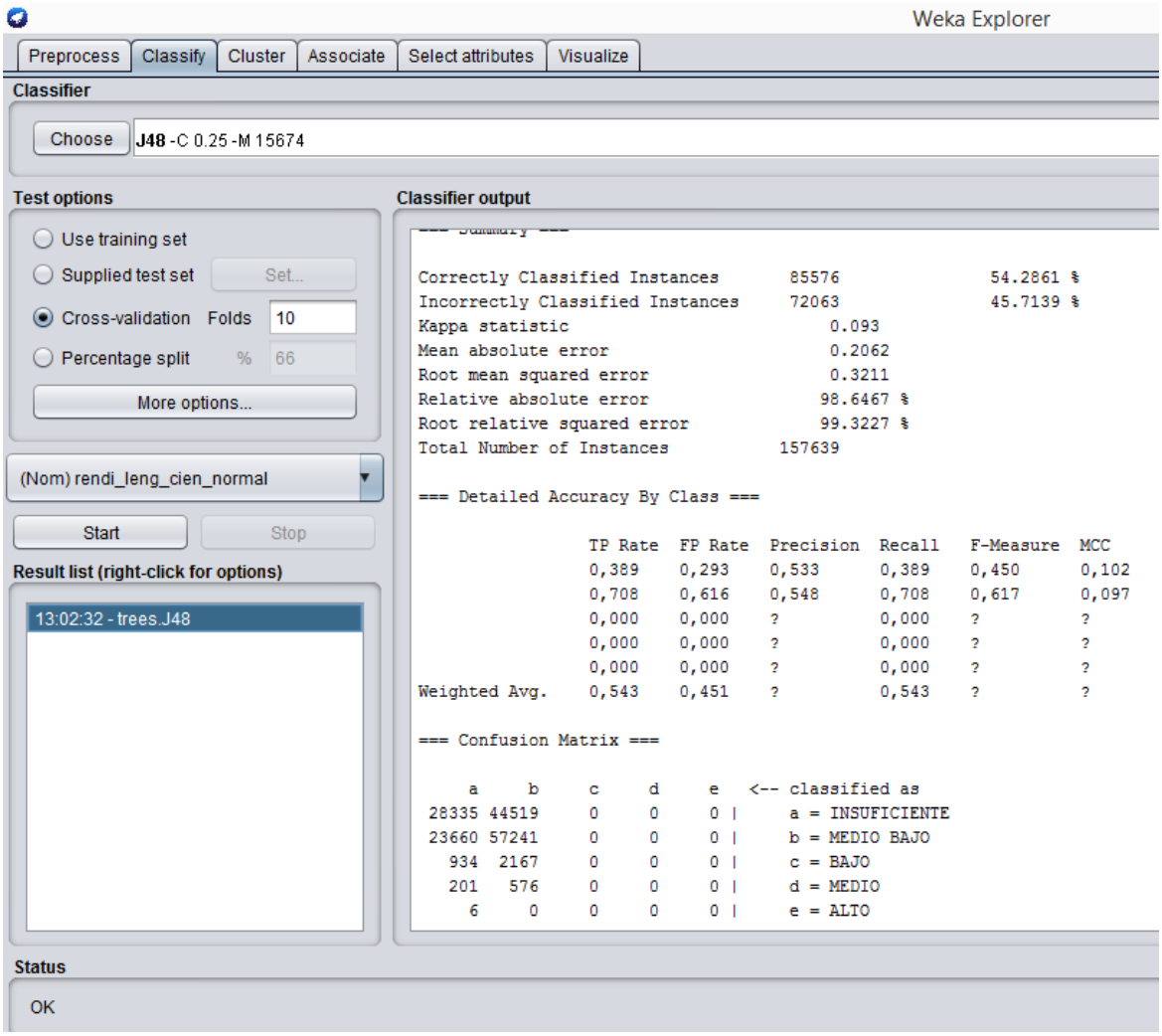

Figura 5. Clasificación con software WEKA.

- **Clasificación con árboles de decisión:** El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender (Han & Kamber, 2001), (Sattler & Dunemann, 2001), (Timarán & Millán, 2006). La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y sólo una hoja, asignando una única clase a la predicción (Hernández & Lorente, 2009).

El algoritmo utilizado para obtener el modelo de clasificación con árboles de decisión es J48 el cual se encuentra incorporado en el programa; este se basa en la utilización del criterio ratio de ganancia (*gain ratio*). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido (Hernández & Lorente, 2009).

El parámetro más importante que se debe tener en cuenta para la poda es el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños (García & Álvarez, 2010). Otra forma de variar el tamaño del árbol es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol, que depende del número absoluto de instancias en el conjunto de datos de partida (Hall, Frank, & Witten, 2011).

Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba.

Existen diferentes medidas de evaluación del clasificador en WEKA a saber:

a. **Usar el conjunto de datos de entrenamiento (*Use training set*):** se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (está técnica puede ser muy buena para ese conjunto de datos, pero puede ser poco precisa para nuevos datos).

b. **Proveer un conjunto de datos de prueba (*Supplied test set*):** se emplea un conjunto de datos para entrenar y otro conjunto independiente al universo de los datos con los que se está trabajando para prueba (corriendo el riesgo que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplearon para entrenar el modelo).

c. **Porcentaje de Partición (*Percentage Split*):** se emplea un % aleatorio de datos para entrenar y otro % para probar, este método difiere del anterior en que ambos conjuntos

pertenecen al universo de datos con el que se está trabajando por lo que se elimina el riesgo que corre el anterior.

d. Validación cruzada (*Cross validation*): Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición (Hernández, Ramírez y Ferri, 2004). Para este caso particular se utiliza el método de evaluación validación cruzada con n pliegues (n-fold cross validation). Está es la opción por defecto y la más comúnmente utilizada. Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (*folds)* de forma aleatoria. El número de subconjuntos se puede introducir en el campo *Folds*. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-1 (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de la n errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (*10-fold cross validation*) (Hernández et al., 2004).

Por otra parte, es bastante sencillo evaluar o estimar el coste de un clasificador para un determinado conjunto de ejemplos si se dispone de la matriz de confusión. La matriz de confusión (*Confusion Matrix*) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1…n constituyen el número de instancias que realmente pertenecen a la clase i. similarmente la sumatoria de los ejemplos o registros en cada columna j, j = 1…n son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y el resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra) (Fernández, 2009). Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba.

Teniendo en cuenta los parámetros de evaluación anteriores y los repositorios de datos descritos en la tabla 35 se procedió a construir los diferentes árboles de decisión con el algoritmo J48. Se escogió como clase el rendi_leng_mate_normal cabe resaltar que esto se realizó con cada repositorio, en cada uno se tuvo en cuenta el rendimiento entre las dos competencias que presenta el mismo, para evaluar la calidad del modelo y su validez se escogió el método de validación cruzada y específicamente la validación cruzada con 10 pliegues por los mejores resultados que se obtienen. En la Figura 6 se muestra está configuración con WEKA.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas se configuraron cuatro valores para el factor de confianza C en 25%, 20%, 15% y 10% manteniendo constante el factor M en 1%. Además, se aplicó un proceso de pos poda para dejar las ramas y por ende las reglas más representativas, que son aquellas que sobrepasan un mínimo soporte del 1% y una confianza del 55%. Finalmente se escoge el árbol para cada competencia con la mayor exactitud de clasificación y que sea fácil de interpretar.

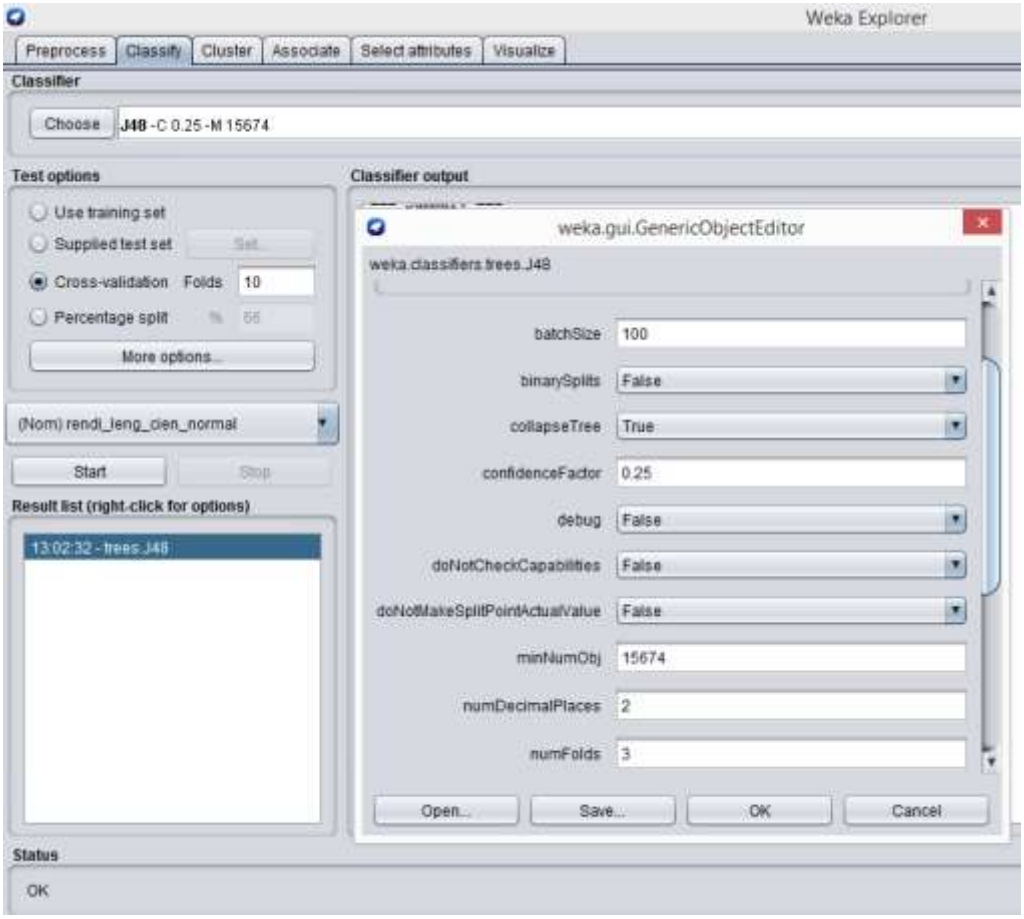

Figura 6. Configuración software WEKA.

4.6.1 Descubrimiento de patrones de desempeño lenguaje y ciencias naturales

Para la construcción del árbol de decisión en el descubrimiento de patrones de desempeño en las competencias de lenguaje y ciencias naturales de los estudiantes de todo el país que presentaron las pruebas Saber 9 entre los años 2014 al 2016, se utilizó el conjunto de datos a_leng_cien_final_final.

El mejor árbol fue el construido con los parámetros M=1576 (1%) y C=0.25 para la pre poda y con confianza mayor o igual al 56,92 % y soporte mayor o igual al 1% para la postpoda.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

sector_cuali = oficial
|   jornada = M
|   |   nivel = 1
|   |   |   zona_cuali = urbana
|   |   |   |   zona_geo = Pacífico: MEDIO BAJO (1983.0/827.0)
|   |   |   |   zona_geo = Caribe: MEDIO BAJO (4822.0/1315.0)
|   |   |   zona_cuali = rural: MEDIO BAJO (21719.0/9660.0)
|   |   nivel = 2
|   |   |   zona_geo = Pacífico: MEDIO BAJO (13395.0/6299.0)
|   |   |   zona_geo = Caribe
|   |   |   |   zona_cuali = urbana: INSUFICIENTE (13038.0/6361.0)
|   |   |   |   zona_cuali = rural: MEDIO BAJO (2423.0/991.0)
|   |   |   zona_geo = Eje Cafetero: MEDIO BAJO (5784.0/2617.0)
|   |   |   zona_geo = Centro Oriente: INSUFICIENTE (9657.0/4349.0)
|   |   |   zona_geo = Centro sur: MEDIO BAJO (6217.0/2828.0)
|   |   |   zona_geo = Llano: MEDIO BAJO (5714.0/2738.0)
|   |   nivel = 3
|   |   |   zona_geo = Pacífico: MEDIO BAJO (2606.0/1003.0)
|   |   |   zona_geo = Centro Oriente: INSUFICIENTE (2946.0/1409.0)
|   |   nivel = 4: MEDIO BAJO (2238.0/1024.0)
|   jornada = C
|   |   zona_geo = Pacífico: MEDIO BAJO (2852.0/1152.0)
|   |   zona_geo = Eje Cafetero
|   |   |   nivel = 2: INSUFICIENTE (5524.0/2628.0)
|   |   |   nivel = 3: MEDIO BAJO (1632.0/671.0)
|   |   zona_geo = Centro Oriente: INSUFICIENTE (11199.0/3983.0)
|   jornada = T
|   |   nivel = 1: MEDIO BAJO (3189.0/1230.0)
|   |   nivel = 2
|   |   |   zona_geo = Caribe: MEDIO BAJO (3685.0/1395.0)
|   |   |   zona_geo = Eje Cafetero: MEDIO BAJO (2723.0/843.0)
|   |   nivel = 3: MEDIO BAJO (3299.0/1093.0)
sector_cuali = no oficial: INSUFICIENTE (8705.0/2941.0)

Number of Leaves  :     45

Size of the tree :  58
```

Figura 7. Árbol para lenguaje - ciencias

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       89729               56.9206 %
Incorrectly Classified Instances     67910               43.0794 %
Kappa statistic                          0.15
Mean absolute error                      0.2007
Root mean squared error                  0.3169
Relative absolute error                 96.0201 %
Root relative squared error             98.0064 %
Total Number of Instances           157639

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 34782 38067     5     0     0 |     a = INSUFICIENTE
 25954 54945     2     0     0 |     b = MEDIO BAJO
```

Figura 8. Precisión y matriz de confusión para las competencias de lenguaje y ciencias.

4.6.2 Descubrimiento de patrones de desempeño lenguaje competencias ciudadanas

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en las competencias de lenguaje y competencias ciudadanas de los estudiantes de todo el país que presentaron las pruebas Saber 9 entre los años 2014 al 2016, se utilizó el conjunto de datos a_leng_comp_final_final.

El mejor árbol fue el construido con los parámetros M=350 (0,5%) y C=0.6 para la pre poda y con confianza mayor o igual al 77,46 % y soporte mayor o igual al 0,5 % para la postpoda.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

zona_cuali = urbana
|   nivel = 1: MEDIO BAJO (7035.0/1585.0)
|   nivel = 2: MEDIO BAJO (19916.0/4144.0)
|   nivel = 3
|   |   zona_geo = Centro Oriente
|   |   |   jornada = M: MEDIO BAJO (2228.0/576.0)
|   |   zona_geo = Eje Cafetero: MEDIO BAJO (3880.0/896.0)
|   |   zona_geo = Caribe: MEDIO BAJO (1113.0/236.0)
|   |   zona_geo = Pacífico: MEDIO BAJO (2592.0/605.0)
|   |   zona_geo = Centro sur: MEDIO BAJO (1764.0/317.0)
|   |   zona_geo = Llano: MEDIO BAJO (1019.0/79.0)
|   nivel = 4: MEDIO BAJO (5428.0/1003.0)
zona_cuali = rural
|   zona_geo = Centro Oriente: MEDIO BAJO (4131.0/923.0)
|   zona_geo = Eje Cafetero
|   |   jornada = C: MEDIO BAJO (2188.0/532.0)
|   |   jornada = M: MEDIO BAJO (867.0/228.0)
|   zona_geo = Caribe: MEDIO BAJO (6831.0/1393.0)
|   zona_geo = Pacífico: MEDIO BAJO (5535.0/1316.0)
|   zona_geo = Centro sur: MEDIO BAJO (2464.0/449.0)
|   zona_geo = Llano: MEDIO BAJO (1046.0/176.0)

Number of Leaves  :     22

Size of the tree :  29
```

Figura 9.Mejor árbol para las competencias de lenguaje y competencias ciudadanas.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       55108               78.5395 %
Incorrectly Classified Instances     15058               21.4605 %
Kappa statistic                          0.0194
Mean absolute error                      0.1414
Root mean squared error                  0.2659
Relative absolute error                 98.4841 %
Root relative squared error             99.257  %
Total Number of Instances            70166

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 54914    79     0     0     0 |     a = MEDIO BAJO
 10668   194     0     0     0 |     b = BAJO
```

Figura 10. Precisión del árbol y su matriz de confusión para las competencias de lenguaje y competencias ciudadanas.

4.6.3 Descubrimiento de patrones de desempeño lenguaje y matemáticas

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en las competencias de lenguaje y matemáticas de los estudiantes de todo el país que presentaron las pruebas Saber 9 entre los años 2014 al 2016, se utilizó el conjunto de datos a_leng_mate_final_final.

El mejor árbol fue el construido con los parámetros M=2700 (1%) y C=0.25 para la pre poda y con confianza mayor o igual al 55% y soporte mayor o igual al 1% para la postpoda.

```
J48 pruned tree
------------------

sector_cuali = oficial
|   jornada = M
|   |   nivel = 1
|   |   |   zona_cuali = urbana
|   |   |   |   zona_geo = Caribe: MEDIO BAJO (8561.0/3651.0)
|   |   nivel = 2
|   |   |   zona_geo = Caribe
|   |   |   |   zona_cuali = rural: MEDIO BAJO (3024.0/1175.0)
|   |   |   |   zona_cuali = urbana: INSUFICIENTE (17181.0/8108.0)
|   |   |   zona_geo = Centro Oriente
|   |   |   |   sexo_cuali = M: INSUFICIENTE (5595.0/2620.0)
|   |   |   |   sexo_cuali = F: INSUFICIENTE (6515.0/2927.0)
|   |   |   zona_geo = Eje Cafetero: MEDIO BAJO (7268.0/3118.0)
|   |   |   zona_geo = Llano: INSUFICIENTE (7728.0/3421.0)
|   |   |   zona_geo = Centro sur
|   |   |   |   zona_cuali = urbana: MEDIO BAJO (7086.0/3414.0)
|   |   |   zona_geo = Centro Oriente
|   |   |   |   sexo_cuali = M: MEDIO BAJO (2282.0/1082.0)
|   |   |   |   sexo_cuali = F: INSUFICIENTE (2835.0/1386.0)
|   |   |   zona_geo = Pacífico: MEDIO BAJO (4913.0/2270.0)
|   |   |   zona_geo = Centro sur: INSUFICIENTE (2520.0/1088.0)
|   |   nivel = 4
|   |   |   sexo_cuali = F: MEDIO BAJO (2459.0/1159.0)
|   jornada = C
|   |   zona_geo = Centro Oriente: INSUFICIENTE (15685.0/6165.0)
|   |   zona_geo = Pacífico: MEDIO BAJO (4161.0/1535.0)
|   |   zona_geo = Eje Cafetero
|   |   |   zona_cuali = rural
|   |   |   |   nivel = 1: INSUFICIENTE (3754.0/1862.0)
|   |   |   |   nivel = 2: INSUFICIENTE (2490.0/1120.0)
|   |   |   zona_cuali = urbana: MEDIO BAJO (7123.0/3009.0)
|   |   zona_geo = Centro sur
|   |   |   nivel = 2: INSUFICIENTE (2547.0/1203.0)
|   jornada = T: MEDIO BAJO (25250.0/10100.0)
sector_cuali = no oficial
|   nivel = 3: INSUFICIENTE (3076.0/969.0)
|   nivel = 4: INSUFICIENTE (8382.0/2759.0)

Number of Leaves  :     55
```

Figura 11 Mejor árbol para las competencias de lenguaje y matemáticas.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      126702               55.7653 %
Incorrectly Classified Instances    100504               44.2347 %
Kappa statistic                          0.1391
Mean absolute error                      0.2048
Root mean squared error                  0.3201
Relative absolute error                 96.9926 %
Root relative squared error             98.4975 %
Total Number of Instances           227206

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 70131 41600     0     0     0 |     a = MEDIO BAJO
 52303 56571     0     0     0 |     b = INSUFICIENTE
```

Figura 12. Precisión del árbol y su matriz de confusión para las competencias de lenguaje y matemáticas.

4.6.4 Descubrimiento de patrones de desempeño matemáticas y competencias ciudadanas

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en las competencias de lenguaje y competencias ciudadanas de los estudiantes de todo el país que presentaron las pruebas Saber 9 entre los años 2014 al 2016, se utilizó el conjunto de datos a_mate_comp_final_final.

El mejor árbol fue el construido con los parámetros M= 695 (1%) y C=0.25 para la pre poda y con confianza mayor o igual al 77.32% y soporte mayor o igual al 1% para la postpoda

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

zona_cuali = urbana: MEDIO BAJO (46551.0/10606.0)
zona_cuali = rural
|   zona_geo = Caribe: MEDIO BAJO (6656.0/1372.0)
|   zona_geo = Centro sur: MEDIO BAJO (2464.0/469.0)
|   zona_geo = Centro Oriente: MEDIO BAJO (4072.0/1015.0)
|   zona_geo = Llano: MEDIO BAJO (1041.0/216.0)
|   zona_geo = Pacífico: MEDIO BAJO (5501.0/1057.0)
|   zona_geo = Eje Cafetero
|   |   jornada = M: MEDIO BAJO (842.0/303.0)
|   |   jornada = C: MEDIO BAJO (2095.0/646.0)

Number of Leaves  :     9

Size of the tree :  12
```

Figura 13. Mejor árbol para las competencias matemáticas y competencias ciudadanas.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53737               77.3239 %
Incorrectly Classified Instances     15759               22.6761 %
Kappa statistic                          0.0182
Mean absolute error                      0.145
Root mean squared error                  0.2693
Relative absolute error                 98.7108 %
Root relative squared error             99.3694 %
Total Number of Instances            69496


=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
   199 13166     0     0     0 |     a = BAJO
    69 53538     0     0     0 |     b = MEDIO BAJO
```

Figura 14. Precisión del árbol y su matriz de confusión para las competencias de matemáticas y competencias ciudadanas.

4.6.5 Descubrimiento de patrones de desempeño Matemáticas y ciencias Naturales

Para estás competencias se optó por utilizar reglas de asociación puesto que los arboles de decisión evidenciaban un solo nodo por la forma que tienen los datos donde un 99% de los datos están clasificados en insuficiente; se establecieron los parámetros del soporte y la confianza así la confianza mínima es 100% (1) y un soporte de 0.1 y un numero de reglas de 1000 y se tienen en cuenta aquellas donde el atributo rendi_mate_cien este como consecuente de la regla y sean por lo menos tres antecedentes. Los resultados al ejecutar el algoritmo a priori se presentan en la Figura 15 y algunas reglas en la Figura 16

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 1000 -T 0 -C 1.0 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -A -c 8
Relation:     a_mate_cien_final_final-weka.filters.unsupervised.attribute.Remove-R2-5,8,12,14-15-
Instances:    156739
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_cien_normal
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (15674 instances)
Minimum metric <confidence>: 1
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 17

Size of set of large itemsets L(2): 65

Size of set of large itemsets L(3): 94

Size of set of large itemsets L(4): 59

Size of set of large itemsets L(5): 16

Size of set of large itemsets L(6): 2

Best rules found:
```

Figura 15. Parámetros de ejecución del algoritmo Apriori para las competencias de matemáticas y ciencias naturales.

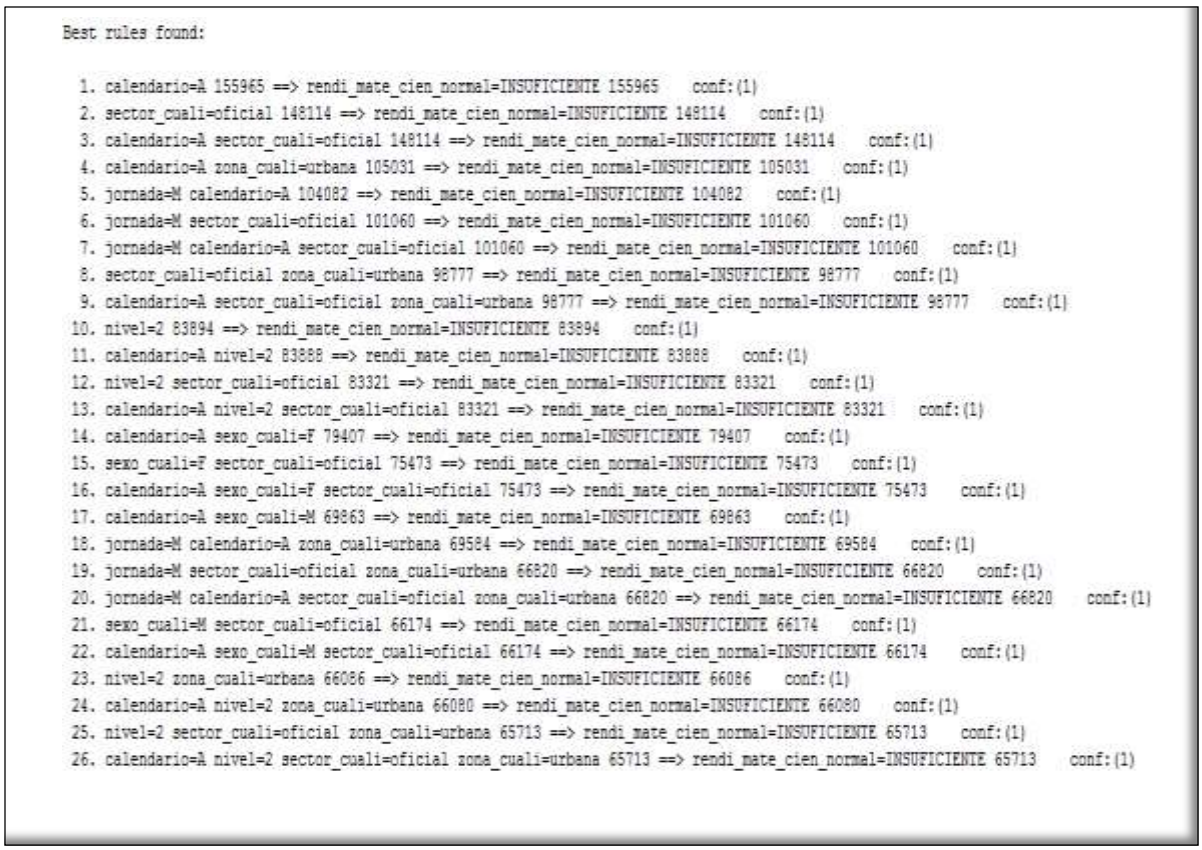

Figura 16. Reglas generadas con el algoritmo Apriori para las competencias matemáticas y ciencias naturales.

4.7 Evaluación

En está fase se evaluaron los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. La evaluación e interpretación de los patrones descubiertos se describe en el capítulo de resultados.

4.8 Implementación

En está fase, el conocimiento descubierto se incorporará al existente y podrá ser utilizado en la toma de decisiones para las instituciones gubernamentales que velan por la calidad de la educación media en Colombia. Una vez, dichas instituciones intervengan en los factores asociados al desempeño académico en las pruebas Saber 9, será posible analizar los resultados y determinar sus efectos.

5. INTERPRETACIÓN Y DISCUSIÓN DE RESULTADOS

5.1 Interpretación

En está sección se evalúan los modelos de clasificación basados en árboles de decisión y las reglas obtenidas con la clasificación basada en asociación por cada competencia y se interpretan los resultados obtenidos en la etapa de modelamiento.

5.1.1 Evaluación e interpretación de resultados para las competencias lenguaje y ciencias naturales

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos a_leng_cien_final_final (ver ***Figura 7***), en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales de 157640 estudiantes que presentaron las pruebas saber 9 entre los años 2014 y 2016 de todo el país en la competencias genéricas de lenguaje y ciencias, donde se escogió el atributo *rendi_leng_cien_normal* como clase, se puede observar que este clasifica correctamente a 85576 instancias, que corresponde a un porcentaje de precisión del 54,2% y 72063 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 45.7% (ver en la Figura 7).

Teniendo en cuenta la distribución de los valores del atributo clase *rendi_leng_cien_normal* del repositorio *a_leng_cien_final_final* que es de 72849 registros para el valor "INSUFICIENTE" y 80889 registros para el valor "MEDIO BAJO" y evaluando el modelo con la matriz de confusión de la Figura 8, este clasifica correctamente a 34782 casos de estudiantes cuyos resultados en las competencias de lenguaje y ciencias están insuficiente y a 54945 encontrándose en medio bajo. Por otra parte, clasifica incorrectamente a 38067 casos de estudiantes cuyos resultados en las competencias de lenguaje y ciencias están insuficiente y 29954 casos que están medio bajo. Esto significa que el modelo clasifica correctamente al 47.7 % de los estudiantes que están insuficiente en las competencias de lenguaje y ciencias y al 67,9 % de las estudiantes que están en medio bajo en está s competencias.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver Figura 7) los patrones más representativos descubiertos, teniendo en cuenta que superen un soporte mínimo del 1% y una confianza mínima del 55,9 %, son:

Regla 1: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 1 y a la zona urbana y además la zona geográfica es Pacifico entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.25 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 58,29 % de los 1983 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.42 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 2: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 1 y a la zona urbana y además la zona geográfica es Caribe entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 3.05 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 72.7 % de los 4822 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 4.33 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 3: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 1 y a la zona rural entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 13.77 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55.5 % de los 21719 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 14.9 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 4: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Pacifico

entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 8.49 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52,9 % de los 13395 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 8.77 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 5: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Caribe y además a la zona urbana entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en INSUFICIENTE. El 8.27 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 51.21 % de los 13038 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 9.16 % de los 72849 estudiantes se encuentran en INSUFICIENTE.

Regla 6: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Caribe y además a la zona rural entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.53 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 61.1 % de los 5784 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.77 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 7: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Eje Cafetero entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 3.66 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 54.7 % de los 2423 de los

estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.91 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 8: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Eje Cafetero entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en INSUFICIENTE. El 6.12 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 54.9 % de los 9657 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 7.28 % de los 72849 estudiantes se encuentran en INSUFICIENTE.

Regla 9: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Centro Sur entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 3.94 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55.9 % de los 6217 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 4.18 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 10: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 2 y a la zona geográfica Llano entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 3.62 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.08 % de los 5714 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.67 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 11: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 3 y a la zona geográfica Pacifico entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.65 % de los

157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 61.51 % de los 2606 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.98 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 12: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 3 y a la zona geográfica Centro Oriente entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en INSUFICIENTE. El 1.86 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.17 % de los 2946 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.89 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 13: si los estudiantes están en una institución del sector oficial en la jornada de la mañana y la institución pertenece al nivel socioeconómico 4 entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.41 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 54.24 % de los 2238 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.34 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 14: si los estudiantes están en una institución del sector oficial en la jornada COMPLETA y la institución pertenece a la zona geográfica Pacifico entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.8 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 59.6 % de los 2852 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.1 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 15: si los estudiantes están en una institución del sector oficial en la jornada COMPLETA y la institución pertenece a la zona geográfica Eje Cafetero y además el nivel socioeconómico de la institución es 2 entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en INSUFICIENTE. El 3.5 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.42 % de los 5524 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.97 % de los 72849 estudiantes se encuentran en INSUFICIENTE.

Regla 16: si los estudiantes están en una institución del sector oficial en la jornada completa y la institución pertenece a la zona geográfica Eje Cafetero y además el nivel socioeconómico de la institución es 3 entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.03 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 58.8 % de los 1632 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.18 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 17: si los estudiantes están en una institución del sector oficial en la jornada completa y la institución pertenece a la zona geográfica Centro Oriente entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en INSUFICIENTE. El 7.1 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 64.43 % de los 11199 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 9.9 % de los 72849 estudiantes se encuentran en INSUFICIENTE.

Regla 18: si los estudiantes están en una institución del sector oficial en la jornada de la tarde y la institución pertenece al nivel socioeconómico 1 entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 2.02 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se

clasifican de está manera. El 61.42 % de los 3189 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.42 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 19: si los estudiantes están en una institución del sector oficial en la jornada de la tarde y la institución pertenece al nivel socioeconómico 2 y además a la zona geográfica Caribe entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 2.33 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 62.14 % de los 3685 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.83 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 20: si los estudiantes están en una institución del sector oficial en la jornada de la tarde y la institución pertenece al nivel socioeconómico 2 y además a la zona geográfica Eje Cafetero entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.72 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 69.04 % de los 2723 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.32 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 21: si los estudiantes están en una institución del sector oficial en la jornada de la tarde y la institución pertenece al nivel socioeconómico 3 entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 2.09 % de los 157639 estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 66.86 % de los 3299 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.72 % de los 80899 estudiantes se encuentran en MEDIO BAJO.

Regla 22: si los estudiantes están en una institución del sector no oficial, entonces el desempeño académico de estos estudiantes en las competencias de lenguaje y ciencias tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 5.52 % de los 157639

estudiantes analizados en las competencias de lenguaje y ciencias en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 62.21 % de los 8705 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 7.91 % de los 72849 estudiantes se encuentran en INSUFICIENTE.

5.1.2 Evaluación e interpretación de resultados para las competencias lenguaje y competencias ciudadanas

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *a_leng_comp_final_final* (ver Figura 9), en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales de 70166 estudiantes que presentaron las pruebas saber 9 entre los años 2014 y 2016 de todo el país en la competencias genéricas de lenguaje y competencias ciudadanas, donde se escogió el atributo *rendi_leng_comp_normal* como clase, se puede observar que este, clasifica correctamente a 55108 instancias, que corresponde a un porcentaje de precisión del 78.53% y 15058 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 21.46% (ver Figura 10).

Teniendo en cuenta la distribución de los valores del atributo clase *rendi_leng_comp_normal* del repositorio a_leng_comp_final_final que es de 54993 registros para el valor “MEDIO BAJO” y 10862 registros para el valor “BAJO” y evaluando el modelo con la matriz de confusión de la Figura 10, este clasifica correctamente a 54914 casos de estudiantes cuyos resultados en las competencias de lenguaje y ciencias están medio bajo y a 194 casos que están bajo. Por otra parte, clasifica incorrectamente a 79 casos de estudiantes cuyos resultados en las competencias de lenguaje y ciencias están medio bajo y 10668 casos que están bajo. Esto significa que el modelo clasifica correctamente al 91.5 % de los estudiantes que están medio bajo en las competencias de lenguaje y competencias ciudadanas y al 1.78 % de las estudiantes que están en bajo en está s competencias.

De acuerdo a las reglas de clasificación obtenidas con este modelo (Figura 10), los patrones más representativos descubiertos, teniendo en cuenta que superen un soporte mínimo del 0.6% y una confianza mínima del 78%, son:

Regla 1: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 1 entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 10 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 77.46 % de los 7035 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 9.91 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 2: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 2 entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 28.38 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 79.19 % de los 19916 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 28.68 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 3: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Centro oriente, además pertenece a jornada de la mañana entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.17 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 74.14 % de los 2228 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 4: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Eje Cafetero, entonces los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 5.52 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 76.9 % de los 3880 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 5.43 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 5: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Caribe entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.58 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 78.79 % de los 1113 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.5 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 6: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Pacifico entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.69 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 76.6 % de los 2592 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.61 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 7: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Centro Sur entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 2.5 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 82 % de los 1764 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.63 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 8: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 3 y la zona geográfica Llano entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.45 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de

está manera. El 92 % de los 1019 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.7 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 9: si la institución se encuentra en la zona urbana y su nivel socioeconómico es 4 entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 7.7 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 81.5 % de los 5428 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 8 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 10: si la institución se encuentra en la zona rural y además a la zona geográfica Centro Oriente entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 5.8 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 77.6 % de los 4131 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 5.8 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 11: si la institución se encuentra en la zona rural y además a la zona geográfica Eje Cafetero y la jornada es completa entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.11 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 76.8 % de los 2188 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 12: si la institución se encuentra en la zona rural y además a la zona geográfica Eje Cafetero y la jornada de la institución es en la mañana entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.23 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 73.7 % de los 867 de los estudiantes analizados que se

clasifican así, son correctamente clasificados y el 1.16 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 13: si la institución se encuentra en la zona rural y además a la zona geográfica Caribe entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 9.7 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 79.6 % de los 6831 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 9.8 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 14: si la institución se encuentra en la zona rural y además a la zona geográfica Pacifico entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 7.8 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 76.2 % de los 5535 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 7.6 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 15: si la institución se encuentra en la zona rural y además a la zona geográfica Centro Sur entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.5 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 81.7 % de los 2464 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.6 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

Regla 15: si la institución se encuentra en la zona rural y además a la zona geográfica Llano entonces; los estudiantes presentan la prueba en las competencias de lenguaje y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.46 % de los 70166 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 83 % de los 1046 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.5 % de los 54993 estudiantes se encuentran en MEDIO BAJO.

5.1.3 Evaluación e interpretación de resultados para las competencias lenguaje y matemáticas

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *a_leng_mate_final_final* (ver Figura 11), en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales de 227296 estudiantes que presentaron las pruebas saber 9 entre los años 2014 y 2016 de todo el país en la competencias genéricas de lenguaje y matemáticas, donde se escogió el atributo *rendi_leng_mate_normal* como clase, se puede observar que este clasifica correctamente a 125702 instancias, que corresponde a un porcentaje de precisión del 55.76% y 100504 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 44.23% (ver **Figura 11**).

Teniendo en cuenta la distribución de los valores del atributo clase *rendi_leng_mate_normal* del repositorio *a_leng_mate_final_final* que es de 111731 registros para el valor "MEDIO BAJO" y 108874 registros para el valor "INSUFICIENTE" y evaluando el modelo con la matriz de confusión de la Figura 12, este clasifica correctamente a 70131 casos de estudiantes cuyos resultados en las competencias de lenguaje y matemáticas están MEDIO BAJO y a 56571 casos que están INSUFICIENTE. Por otra parte, clasifica incorrectamente a 41600 casos de estudiantes cuyos resultados en las competencias de lenguaje y matemáticas están MEDIO BAJO y 52303 casos que están INSUFICIENTE. Esto significa que el modelo clasifica correctamente al 62.7 % de los estudiantes que están medio bajo en las competencias de lenguaje y matemáticas y al 51.9% de las estudiantes que están en insuficiente en está s competencias.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver Figura 13), los patrones más representativos descubiertos, teniendo en cuenta que superen un soporte mínimo del 1% y una confianza mínima del 55.76%, son:

Regla 1: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es uno y además la zona de la institución es urbana y se encuentra en la zona geográfica Caribe entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en

MEDIO BAJO. El 3.76% de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 57.35 % de los 8561 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 4,39 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 2: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y además la zona de la institución es rural y se encuentra en la zona geográfica Caribe entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.33 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 57.35 % de los 3041 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.65 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 3: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y además la zona de la institución es urbana y se encuentra en la zona geográfica Caribe entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 7.56 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.8 % de los 17181 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 8.3 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 4: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y los estudiantes que presentan la prueba son de género masculino y la institución se encuentra en la zona geográfica Centro Oriente entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 2.46 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 53.17 % de los 5595 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.64 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 5: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y los estudiantes que presentan la prueba son de género femenino y la institución se encuentra en la zona geográfica Centro Oriente entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 2.86 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55 % de los 6515 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.98 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 6: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y la institución se encuentra en la zona geográfica Eje Cafetero entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.18 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55.9 % de los 7268 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.71 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 7: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y la institución se encuentra en la zona geográfica Llano entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 3.18 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55.7 % de los 7768 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.9 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 8: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y la institución se encuentra en la zona geográfica Centro Sur y además la institución se encuentra en una zona urbana entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de

está r clasificados en MEDIO BAJO. El 3.11 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 51.8 % de los 7086 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.21 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 9: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es dos y la institución se encuentra en la zona geográfica Centro Sur y además la institución se encuentra en una zona urbana entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.11 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 51.8 % de los 7086 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.21 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 10: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es tres y la institución se encuentra en la zona geográfica Centro Oriente y el género del estudiante es masculino entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.12 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.8 % de los 2282 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.09 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 11: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es tres y la institución se encuentra en la zona geográfica Centro Oriente y el género del estudiante es femenino entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.24 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 51.2 % de los 2835 de los estudiantes analizados que se clasifican

así, son correctamente clasificados y el 1.33 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 12: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es tres y la institución se encuentra en la zona geográfica Pacifico entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 2.16 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 53.7 % de los 4913 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.36 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 13: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es tres y la institución se encuentra en la zona geográfica Centro Sur entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.2 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 56.8 % de los 2520 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.32% de los 111731 estudiantes se encuentran en INSUFICIENTE.

Regla 14: Si la institución está adscrita al sector oficial y pertenece al a jornada de la mañana y el nivel socioeconómico es cuatro y los estudiantes que presentan la prueba son de género femenino entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.1 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.8 % de los 2459 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.17 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 15: Si la institución está adscrita al sector oficial y pertenece a la jornada completa y a la zona geográfica Centro Oriente entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 6.9 % de los 227206 estudiantes analizados en las competencias de

lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 60.69 % de los 15685 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 8.74 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 16: Si la institución está adscrita al sector oficial y pertenece a jornada completa y a la zona geográfica Pacifico entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.83 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 63.1 % de los 4161 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.35 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 17: Si la institución está adscrita al sector oficial y pertenece a jornada completa y a la zona geográfica Eje Cafetero y la institución se encuentra en zona rural y pertenece al nivel socioeconómico uno entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.65 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 50.4 % de los 3754 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.73 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 18: Si la institución está adscrita al sector oficial y pertenece a la jornada completa y a la zona geográfica Eje Cafetero y la institución se encuentra en zona rural y pertenece al nivel socioeconómico dos entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.1 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 55 % de los 2490 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.25 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 19: Si la institución está adscrita al sector oficial y pertenece a la jornada completa y a la zona geográfica Eje Cafetero y la institución se encuentra en zona urbana y pertenece al nivel socioeconómico dos entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.13 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 57.7 % de los 2490 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.6 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 20: Si la institución está adscrita al sector oficial y pertenece a jornada completa y a la zona geográfica Centro Sur y la institución pertenece al nivel socioeconómico dos entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.12 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 52.7% de los 2547 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.23 % de los 111731 estudiantes se encuentran en INSUFICIENTE.

Regla 21: Si la institución está adscrita al sector oficial y pertenece a la jornada de la tarde entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 11.1 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 60% de los 25250 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 13.5 % de los 111731 estudiantes se encuentran en MEDIO BAJO.

Regla 22: Si la institución está adscrita al sector no oficial y pertenece al nivel socio económico tres entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 1.35 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 68.4% de los 3076 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.93 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

Regla 23: Si la institución está adscrita al sector no oficial y pertenece al nivel socio económico cuatro entonces; el desempeño académico de estos estudiantes en las competencias de lenguaje y matemáticas tiene mayor probabilidad de está r clasificados en INSUFICIENTE. El 3.68 % de los 227206 estudiantes analizados en las competencias de lenguaje y matemáticas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 67% de los 8382 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 5.16 % de los 108874 estudiantes se encuentran en INSUFICIENTE.

5.1.4 Evaluación e interpretación de resultados para las competencias de matemáticas y competencias ciudadanas

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *a_mate_coomp_final_final* (ver ***Figura 13***), en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales de 69496 estudiantes que presentaron las pruebas saber 9 entre los años 2014 y 2016 de todo el país en la competencias genéricas de matemáticas y Competencias ciudadanas, donde se escogió Figura 13el atributo *rendi_mate_comp_normal* como clase, se puede observar que este clasifica correctamente a 53737 instancias, que corresponde a un porcentaje de precisión del 77.32% y 15759 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 22.67%, (ver **Figura 13**).

Teniendo en cuenta la distribución de los valores del atributo clase *rendi_mate_comp_normal* del repositorio a_mate_comp_final_final que es de 13365 registros para el valor “BAJO” y 53607 registros para el valor “MEDIO BAJO” y evaluando el modelo con la matriz de confusión de la Figura 14, este clasifica correctamente a 199 casos de estudiantes cuyos resultados en las competencias de matemáticas competencias ciudadanas están BAJO y a 53538 casos que están MEDIO BAJO. Por otra parte, clasifica incorrectamente a 13166 casos de estudiantes cuyos resultados en las competencias de matemáticas y competencias ciudadanas están BAJO y 69 casos que están MEDIO BAJO. Esto significa que el modelo clasifica correctamente al 1.49 % de los estudiantes que están bajo en las competencias de matemáticas y competencias ciudadanas y al 99.2% de las estudiantes que están en medio bajo en está s competencias.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver Figura 14), los patrones más representativos descubiertos, teniendo en cuenta que superen un soporte mínimo del 1% y una confianza mínima del 77.32 %, son:

Regla 1: si la institución pertenece a la zona urbana los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 66.9 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 77.2 % de los 46551 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 67 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 2: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Caribe entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 9.57 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 79.3 % de los 6656 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 9.85 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 3: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Centro Sur entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3.5 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 80.9 % de los 2464 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 3.39 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 4: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Centro Oriente entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 5.85 % de los 69496 estudiantes analizados en las competencias de lenguaje

y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 75 % de los 4072 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 5.7 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 5: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Llano entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de estar clasificados en MEDIO BAJO. El 1.49 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 79.2 % de los 4072 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1.53 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 6: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Pacifico entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 7.91 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 80.7 % de los 5501 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 8.28 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 7: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Eje Cafetero y además la institución preste el servicio en jornada de la mañana entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 1.22 % de los 69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 64 % de los 842 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 1 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

Regla 8: si la institución pertenece a la zona rural y está se ubica en la zona geográfica Eje Cafetero y además la institución preste el servicio en jornada de la mañana entonces; los estudiantes que presentan la prueba en las competencias de matemáticas y competencias ciudadanas tiene mayor probabilidad de está r clasificados en MEDIO BAJO. El 3 % de los

69496 estudiantes analizados en las competencias de lenguaje y competencias ciudadanas en las pruebas saber 9 entre los años 2014 al 2016 se clasifican de está manera. El 69 % de los 2095 de los estudiantes analizados que se clasifican así, son correctamente clasificados y el 2.7 % de los 53607 estudiantes se encuentran en MEDIO BAJO.

5.1.5 Evaluación e interpretación de resultados para las competencias de matemáticas y ciencias naturales

Teniendo en cuenta los datos obtenidos con el conjunto de datos *a_mate_cien_final_final* en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales de 156739 estudiantes que presentaron las pruebas saber 9 entre los años 2014 y 2016 de todo el país en la competencias genéricas de matemáticas y ciencias naturales , se puede observar que un 99% de estos datos están en un rango de Insuficiente y menos del 1% se encuentra en un rango diferente por está razón no se puede considerar realizar un árbol de decisión; así que se realiza otro tipo de análisis como son las reglas de asociación a priori el cual nos brinda algunas reglas de asociación cabe resaltar que todas las reglas nos llevan a clasificarse como insuficiente por las razones antes mencionadas.

A continuación, se describen los resultados obtenidos. En los datos de inicio se solicitan 1000 reglas con un soporte de c de 1 y una confianza M de 0,1; con estos datos el programa arroja (ver Figura 15).

Reglas con un antecedente y un consecuente 17
Reglas con dos antecedentes y un consecuente 65
Reglas con tres antecedentes y un consecuente 95
Reglas con cuatro antecedentes u un consecuente 59
Reglas con cinco antecedentes y un consecuente 16
Reglas con seis antecedentes y un consecuente 2.

Teniendo en cuenta las reglas generadas por el algoritmo A priori con el conjunto de datos para las competencias de matemáticas y ciencias naturales se van analizar algunas reglas con tres antecedentes y como consecuente el atributo *rendi_mate_cien_normal*.

Regla 1: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño de Insuficiente pertenecen a, establecimientos con calendario A, el nivel socioeconómico del mismo dos, y el establecimiento es oficial, todos los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 2: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño de Insuficiente pertenecen a, establecimientos con calendario A, el nivel socioeconómico del mismo dos, y el establecimiento es oficial, todos los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 3: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos con calendario A, son estudiantes de género femenino y el establecimiento es oficial, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 4: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño de Insuficiente pertenecen a, establecimientos oficiales, en la jornada de la mañana, con calendario A y la zona del establecimiento es urbana, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 5: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos de a la jornada de la mañana, sector oficial y la zona del establecimiento es urbana, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 6: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente, pertenecen a, establecimientos con calendario A, los estudiantes son de género masculino y el sector del establecimiento es oficial, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 7: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos de la jornada de la mañana, el nivel socioeconómico es dos y el sector del establecimiento es oficial, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 8: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos de calendario A, sector oficial y a la zona rural, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 9: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos de jornada de la mañana, nivel socioeconómico dos, sector oficial y zona urbana, los estudiantes que presentaron está prueba tienen el mismo patrón.

Regla 10: el 100 % de los estudiantes que presentaron la prueba de matemáticas y ciencias naturales obtuvieron un nivel de desempeño Insuficiente pertenecen a, establecimientos de calendario A, sector oficial y zona geográfica caribe, los estudiantes que presentaron está prueba tienen el mismo patrón.

5.2 Discusión de resultados

El objetivo de está investigación fue descubrir patrones de desempeño académico de las pruebas Saber 9 con técnicas predictivas de minería de datos, a partir de la información almacenada en la base de datos ICFES en el periodo 2014-2016. Para cumplir este objetivo se escogieron la clasificación basada en asociación y la clasificación basada en árboles de decisión y para este caso se utilizó J48. Se analizaron los casos de estudiantes que presentaron dicha prueba.

Con las técnicas aplicadas en la presente investigación como el modelo de clasificación basado en árboles de decisión al igual que el modelo basado en asociación generado con el algoritmo Apriori, muestran que los factores asociados al rendimiento académico de las pruebas saber 9 se encuentran en un nivel de desempeño INSUFICIENTE, Este resultado confirma lo expuesto por la OCDE(2016), porque pese a los esfuerzos que hace el gobierno nacional por promover programas que ayuden al mejoramiento de la calidad de educación, los estudiantes continúan obteniendo puntajes muy bajos en las diferentes competencias que evalúa la OCDE con relación a otros países miembros de este ente.

Además en pruebas internacionales como la PISA los resultados muestran que Colombia obtuvo los mejores resultados en lenguaje seguido por ciencias y por último en matemáticas esto muestra que se ha avanzado, pero aun no supera la media de los países que integran la OCDE; está entidad también evalúa el efecto que tienen sobre el aprendizaje las variables socioeconómicas y culturales, así como las características del sistema escolar y las instituciones educativas. (MEN, Ministerio de Educacion, 2019).

Entre los patrones de desempeño académico de las competencias genéricas de las pruebas saber 9, descubiertos y considerados en está discusión, son los factores asociados relacionados con los planteles educativos; se encuentran articulados con atributos como el sector del establecimiento (oficial y no oficial), en el cual no se encontraron diferencias, es decir el sector de la institución no influye en el desempeño académico en las competencias genéricas porque en los dos sectores se obtuvieron los mismos resultados; lo cual contradice lo expuesto por, (Piñeros & Rodriguez, 1999) y (Gomez & Jaramillo, 2017)quienes muestra la importancia de estos aspectos en el aprendizaje, de hecho las diferencias entre las instituciones privadas y públicas, quienes concluyen que hay un mejor desempeño académico en estudiantes que asisten a una institución privada.

En está línea también están los estudios realizados por Gaviria y Barrientos (2001 a y b), quienes determinaron que las características asociadas a las instituciones educativas como la zona, el sector y la jornada influyen de manera significativa en el rendimiento académico.

Otro factor asociado que se encontró en está investigación es la zona donde se encuentra ubicada la institución educativa, (rural o urbano), porque se puede apreciar que no existe ninguna influencia de este factor puesto que en los dos casos el desempeño académico en las competencias genéricas que evalúa las pruebas saber 9 son iguales, lo cual es opuesto a los estudios realizados por (Gomez & Jaramillo, 2017), donde encontraron que la zona de la institución educativa es de gran incidencia en el desempeño académico teniendo en cuenta que algunas instituciones se encuentran más aisladas que otras y esto dificulta la posibilidad de acceder a mejores recursos.

También un factor asociado es el género del estudiante, el cual en está investigación indican que es irrelevante porque en los resultados obtenidos tanto para hombres como para mujeres son los mismos, lo cual es opuesto a estudios realizados por la UNESCO (2010) donde los resultados indican que el género masculino obtiene mejores desempeños en la competencia genérica de Matemáticas.

Seguramente, una de las causas de este desempeño se debe al cambio de políticas de deducción nacional (decreto 230 de 11 de febrero de 2002) por el cual la evaluación se convirtió muy flexible teniendo en cuenta que, uno de los pilares era la promoción automática la cual establecía que el 95% de los estudiantes debería ser promovido al siguiente grado y perdía el año con más de 3 áreas y por inasistencia continua, además contemplaba que los estudiantes que perdían 1 o 2 materias deberían ser promovidos y al comenzar el nuevo año lectivo debían presentar exámenes de recuperación; claramente esto hizo que la educación en Colombia bajara en calidad.

Como se puede ver en los estudios realizados por (Rojas Rubio, 1992) quien en su investigación “Promoción Automática y Fracaso Escolar en Colombia” muestra como está política educativa fue un fracaso puesto que, con está s políticas se pretendía brindar mejores oportunidades de aprendizaje a los sectores poblacionales donde existe más deficiencia; sin embargo los resultados fueron muy contradictorios frente a las expectativas generadas de deserción y resistencia escolar, bajos niveles académicos entre otros.

De igual manera con el 16 de abril del 2009 con el decreto 1290, el cual reglamenta la promoción y evaluación de estudiantes de primaria, secundaria y media académica en forma autónoma para las instituciones públicas y privadas del país, lo único a tener en cuenta son estos criterios que deben centrarse en competencias y la escala que utilice la institución debe ser validada con la escala nacional, todo esto muestra como la evaluación continua siendo flexible.

Por otra parte, se resalta que las políticas de educación nacional, también establecieron el índice sintético (ISCE) de calidad educativa con el cual en realidad se mide a las instituciones públicas y privadas del país; el ISCE el cual evalúa a las instituciones en cuatro aspectos a saber

El ambiente escolar se centra en un ambiente en el aula y el seguimiento de aprendizaje, el cual toma la información de factores asociados de las pruebas saber 5° y 9°.

El desempeño refleja los puntajes que obtuvieron los estudiantes en lenguaje y matemáticas de las pruebas saber pasadas las cuales tienen un puntaje de 100 a 500 y esto permite saber cómo se encuentran las instituciones con referencia al resto del país y así buscar estrategias que permitan incrementarlos en las próximas pruebas saber.

La eficiencia es la calificación que obtienen las instituciones, corresponden a la tasa de aprobación que tienen las mismas, es decir, entre más estudiantes promovidos mayor puntaje obtiene la institución y viceversa esto condiciona a las instituciones para la promoción de los estudiantes.

El progreso indica cuanto han progresado las instituciones con respecto al año anterior, es decir, se comparan las instituciones consigo mismas. También hay que tener en cuenta que hoy por hoy la cobertura en las instituciones es muy amplia lo cual ha llevado a un hacinamiento en las aulas lo cual no es favorable para la calidad educativa.

Además, si el plantel logra cumplir con sus metas teniendo en cuenta los cuatro componentes, este obtendrá incentivos económicos (salario adicional), eso dependerá del nivel de mejoría de cada colegio. (MEN, Colombia Aprende, 2019)

Así el instrumento implementado por el MEN no garantiza la calidad educativa porque mide Colombia con Colombia sin tener en cuenta que la educación del siglo XXI se hace a nivel mundial, es decir se mide a las instituciones educativas del país de carácter privado como público, además no mide la calidad en términos de puntaje sino en ranking, es decir, la supuesta calidad que evalúa donde puede mejorarse por una caída general en los resultados de otros colegios aunque no necesariamente se haga por una mejora en la calidad de una institución.

Esto se puede evidenciar con un estudio realizado por (Antolinez & Jimenez, 2018), quienes realizaron un análisis de correlaciones para los años 2015 al 2017 de los cuatro componentes que este mide y para cada nivel como son primaria, secundaria y media donde dejan abiertas muchas discusiones acerca de los componentes del ISCE y la medición de la calidad educativa en Colombia.

Se puede apreciar que el propósito ISCE es dar una jerarquía a las instituciones educativas en cuanto a la calidad educativa entregando a las mejores instituciones incentivos con el fin de continuar mejorando, pero a la vez ejercer presión sobre las mismas, además tener en cuenta que, no todas las instituciones cuentan con los mismos recursos, ni estructuras físicas, entre otros elementos que influyen para que las instituciones presten un servicio con calidad a sus estudiantes.

6. CONCLUSIONES Y RECOMENDACIONES

El objetivo de está investigación fue descubrir patrones de desempeño académico en las competencias genéricas de las pruebas Saber 9 que presentaron los estudiantes de todo el país en los periodos 2014 al 2016 utilizando técnicas de minería de datos, a partir de los datos suministrados por el ICFES, de está base de datos se seleccionaron diferentes atributos de los estudiantes. La metodología utilizada para cumplir este objetivo fue CRISP-DM y la técnica de minería de datos aplicada para el descubrimiento de patrones de desempeño académico fue clasificación basada en árboles de decisión y clasificación basada en asociación.

Para el descubrimiento de patrones en cada competencia se tuvo en cuenta aquellos que sobrepasaron un soporte mínimo del 1% del total de casos evaluados y una confianza del 55 % de casos que cumplían el patrón descubierto. El resto de patrones fueron descartados, entre los atributos que forman parte de los patrones descubiertos en las competencias de lenguaje y ciencias naturales, se destacan la jornada, nivel socioeconómico y zona geográfica de los establecimientos como variables importantes asociadas al desempeño académico de los estudiantes que presentaron las pruebas Saber 9.

De acuerdo a los resultados obtenidos algo que sorprende son las instituciones educativas no oficiales en las cuales el nivel de desempeño es insuficiente en está s competencias, debido que, está s instituciones cuentan con mejores infraestructura física y docente.

En las competencias de lenguaje y competencias ciudadanas los atributos que hacen parte de la mayoría de patrones descubiertos son la zona, el nivel socioeconómico y la zona geográfica a la cual pertenecen las instituciones educativas como las variables más importantes asociadas al desempeño académico de los estudiantes que presentaron las pruebas Saber 9.

En las competencias de lenguaje y matemáticas los atributos que más se destacan en la mayoría de los patrones descubiertos son la jornada, el nivel socioeconómico y la zona (urbana, rural) a la cual pertenecen las instituciones educativas como las variables más significativas asociadas al desempeño académico de los estudiantes que presentaron las pruebas Saber 9.

En las competencias de matemáticas y competencias ciudadanas los atributos que más se destacan en la mayoría de los patrones descubiertos son la zona (urbana rural) y zona geográfica a la cual pertenecen las instituciones educativas como las variables más significativas asociadas al desempeño académico de los estudiantes que presentaron las pruebas Saber 9.

En las competencias de matemáticas y ciencias naturales no se utilizó los arboles de decisión puesto que el 99% de los datos están en un nivel de insuficiente y el 1% se encuentra en un nivel diferente por está razón se optó por utilizar el modelo de asociación en el algoritmo A priori el cual también nos muestra que todas las reglas de asociación nos llevan a un nivel de desempeño insuficiente académico de los estudiantes que presentaron las pruebas Saber 9.

En general se puede ver que el desempeño académico en las pruebas saber 9 es insuficiente, lo cual es un panorama desolador y nos hace ver que las políticas nacionales de educación no están funcionando bien por ello se deben implementar estrategias y planes de mejoramiento para fortalecer las instituciones el país en cuanto al acceso de recursos tecnológicos, infraestructura física, docentes idóneos entre otros factores y así mejorar la calidad en la educación media del país.

Se plantea como recomendaciones el complementar este estudio con los resultados de las pruebas saber 9 de los años 2017 y 2018 para evaluar si los malos resultados continúan o no.

Por otra parte se puede hacer otros estudios aplicando otras técnicas de minería de datos como son las descriptivas para analizar qué factores académicos están asociados al rendimiento académico y como se agrupan los estudiantes de acuerdo a este rendimiento

7. ANEXOS

a_leng_cien_final

Clasificador Part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 15764 -C 0.25 -Q 1
Relation:     a_leng_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    157639
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

jornada = M AND
nivel = 2: MEDIO BAJO (56463.0/27970.0)

jornada = M: MEDIO BAJO (48527.0/22331.0)

jornada = C: INSUFICIENTE (34444.0/15350.0)

: MEDIO BAJO (18205.0/6895.0)

Number of Rules  :  4


Time taken to build model: 1.33 seconds
```

```
=== Summary ===

Correctly Classified Instances       85093               53.9797 %
Incorrectly Classified Instances     72546               46.0203 %
Kappa statistic                          0.0758
Mean absolute error                      0.2062
Root mean squared error                  0.3211
Relative absolute error                 98.6186 %
Root relative squared error             99.3084 %
Total Number of Instances           157639

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
               0,262    0,181    0,554      0,262    0,356      0,098
               0,816    0,745    0,536      0,816    0,647      0,085
               0,000    0,000    ?          0,000    ?          ?
               0,000    0,000    ?          0,000    ?          ?
               0,000    0,000    ?          0,000    ?          ?
Weighted Avg.  0,540    0,466    ?          0,540    ?          ?

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 19094 53760     0     0     0 |     a = INSUFICIENTE
 14902 65999     0     0     0 |     b = MEDIO BAJO
   387  2714     0     0     0 |     c = BAJO
    55   722     0     0     0 |     d = MEDIO
     6     0     0     0     0 |     e = ALTO
```

Clasificador Part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 31528 -C 0.25 -Q 1
Relation:     a_leng_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    157639
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

jornada = M: MEDIO BAJO (104990.0/50301.0)

: MEDIO BAJO (52649.0/26437.0)

Number of Rules  :  2


Time taken to build model: 0.83 seconds
```

```
=== Summary ===

Correctly Classified Instances       80901               51.3204 %
Incorrectly Classified Instances     76738               48.6796 %
Kappa statistic                          0
Mean absolute error                      0.2088
Root mean squared error                  0.3231
Relative absolute error                 99.8764 %
Root relative squared error             99.9394 %
Total Number of Instances           157639

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
               0,000     0,000     ?           0,000   ?          ?
               1,000     1,000     0,513       1,000   0,678      ?
               0,000     0,000     ?           0,000   ?          ?
               0,000     0,000     ?           0,000   ?          ?
               0,000     0,000     ?           0,000   ?          ?
Weighted Avg.  0,513     0,513     ?           0,513   ?          ?

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
     0 72854     0     0     0 |     a = INSUFICIENTE
     0 80901     0     0     0 |     b = MEDIO BAJO
     0  3101     0     0     0 |     c = BAJO
     0   777     0     0     0 |     d = MEDIO
     0     6     0     0     0 |     e = ALTO
```

Clasificador Zero R

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 15764
Relation:     a_leng_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    157639
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       80901               51.3204 %
Incorrectly Classified Instances     76738               48.6796 %
Kappa statistic                          0
Mean absolute error                      0.2091
Root mean squared error                  0.3233
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           157639
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 32528
Relation:     a_leng_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    157639
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       80901               51.3204 %
Incorrectly Classified Instances     76738               48.6796 %
Kappa statistic                          0
Mean absolute error                      0.2091
Root mean squared error                  0.3233
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           157639
```

A_leng_comp_final

Clasificador part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 70166 -C 0.5 -Q 1
Relation:     a_leng_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    70166
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: MEDIO BAJO (70166.0/15173.0)

Number of Rules  :  1


Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       54993               78.3756 %
Incorrectly Classified Instances     15173               21.6244 %
Kappa statistic                          0
Mean absolute error                      0.1436
Root mean squared error                  0.2679
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 70166 -C 0.25 -Q 1
Relation:     a_leng_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    70166
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: MEDIO BAJO (70166.0/15173.0)

Number of Rules  :  1


Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       54993               78.3756 %
Incorrectly Classified Instances     15173               21.6244 %
Kappa statistic                          0
Mean absolute error                      0.1436
Root mean squared error                  0.2679
```

Clasificador Zero R

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     a_leng_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    70166
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       54993               78.3756 %
Incorrectly Classified Instances     15173               21.6244 %
Kappa statistic                          0
Mean absolute error                      0.1436
Root mean squared error                  0.2679
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            70166
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 70166
Relation:     a_leng_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    70166
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       54993               78.3756 %
Incorrectly Classified Instances     15173               21.6244 %
Kappa statistic                          0
Mean absolute error                      0.1436
Root mean squared error                  0.2679
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            70166
```

A_leng_mate_final

Clasificador part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 2700 -C 0.25 -Q 1
Relation:     a_leng_mate_final_final_1-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    227206
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_mate_normal
Test mode:    10-fold cross-validation
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 5400 -C 0.25 -Q 1
Relation:     a_leng_mate_final_final_1-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    227206
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_mate_normal
Test mode:    10-fold cross-validation
```

Clasificador Zero R

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     a_leng_mate_final_final_1-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    227206
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_mate_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      111731               49.1761 %
Incorrectly Classified Instances    115475               50.8239 %
Kappa statistic                          0
Mean absolute error                      0.2112
Root mean squared error                  0.3249
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           227206
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 2700
Relation:     a_leng_mate_final_final_1-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    227206
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_leng_mate_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      111731               49.1761 %
Incorrectly Classified Instances    115475               50.8239 %
Kappa statistic                          0
Mean absolute error                      0.2112
Root mean squared error                  0.3249
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           227206
```

A_mate_cien_final

Clasificador part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 15664 -C 0.25 -Q 1
Relation:     a_mate_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    156739
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: INSUFICIENTE (156739.0/266.0)

Number of Rules  :  1


Time taken to build model: 0.97 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      156473               99.8303 %
Incorrectly Classified Instances       266                0.1697 %
Kappa statistic                          0
Mean absolute error                      0.0023
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 31328 -C 0.25 -Q 1
Relation:     a_mate_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    156739
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: INSUFICIENTE (156739.0/266.0)

Number of Rules  :  1


Time taken to build model: 0.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      156473               99.8303 %
Incorrectly Classified Instances       266                0.1697 %
Kappa statistic                          0
Mean absolute error                      0.0023
```

Calsificador Zero R

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     a_mate_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    156739
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: INSUFICIENTE

Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      156473               99.8303 %
Incorrectly Classified Instances       266                0.1697 %
Kappa statistic                          0
Mean absolute error                      0.0023
Root mean squared error                  0.0336
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           156739
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 15664
Relation:     a_mate_cien_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    156739
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_cien_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: INSUFICIENTE

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      156473               99.8303 %
Incorrectly Classified Instances       266                0.1697 %
Kappa statistic                          0
Mean absolute error                      0.0023
Root mean squared error                  0.0336
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           156739
```

A_mate_comp_final

Clasificador Part

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 6950 -C 0.25 -Q 1
Relation:     a_mate_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    69496
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: MEDIO BAJO (69496.0/15889.0)

Number of Rules  :  1


Time taken to build model: 0.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53607               77.1368 %
Incorrectly Classified Instances     15889               22.8632 %
Kappa statistic                          0
Mean absolute error                      0.1469
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.PART -M 13900 -C 0.25 -Q 1
Relation:     a_mate_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    69496
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

PART decision list
------------------

: MEDIO BAJO (69496.0/15889.0)

Number of Rules  :  1


Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53607               77.1368 %
Incorrectly Classified Instances     15889               22.8632 %
Kappa statistic                          0
Mean absolute error                      0.1469
```

Clasificador Zero R

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     a_mate_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    69496
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53607               77.1368 %
Incorrectly Classified Instances     15889               22.8632 %
Kappa statistic                          0
Mean absolute error                      0.1469
Root mean squared error                  0.271
Relative absolute error                100       %
Root relative squared error            100       %
Total Number of Instances            69496
```

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR -batch-size 13900
Relation:     a_mate_comp_final_final-weka.filters.unsupervised.attribute.
Remove-R2-5,8,12,14-15-weka.filters.unsupervised.attribute.
NumericToNominal-Rfirst-last
Instances:    69496
Attributes:   8
              jornada
              calendario
              nivel
              sexo_cuali
              sector_cuali
              zona_cuali
              zona_geo
              rendi_mate_comp_normal
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: MEDIO BAJO

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       53607               77.1368 %
Incorrectly Classified Instances     15889               22.8632 %
Kappa statistic                          0
Mean absolute error                      0.1469
Root mean squared error                  0.271
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            69496
```

8. REFERENCIAS BIBLIOGRAFICAS

(s.f.).

Acosta, ,. A., & Barahona, ,. J. (2018). *Un Mercadode datos para le analisismultidimensional de las pruebas saber 9 de las instituciones educativas de los municipios de la subregion de Obando del depaartamento de Nariño.* Ipiales, Nariño, Colombia.

Antolinez, ,. J., & Jimenez, ,. (2018). *Correlaciones de las componentes de los índices sintéticos de calidad.* Bogota, Colombia.

Ayala Garcia, J. (2015). *Evaluacion externa y calidad de la educacion en Colombia.* Cartagena, Colombia.

Chica, S., Galvis, D., & Ramírez, A. (2010). Determinantes de rendimiento academico en Colombia: pruebas ICFES Saber 11, 2009 ISSN: 0120341X. *Revista Universidad EAFIT, Vol 46, Num. 160 Medellin, Colombia.*

Coleman, J., Campbell, , E., Hobson,, C., McPartland,, J., Mood,, A., & weinfeld,, F. (1996). *Equality of Education Opportunity. National Center for Educational Statistitcs.* Washington. USA.

Erazo, O. (2012). El rendimiento academico, un fenomeno de multiples relaciones y complejidades. *Revista Vanguardia Psicologica*, 144-173.

Fernandez, H. (2005). *Como interpretar la evaluacion Saber. Subdreccion de EStandares y Evauacon. Ministerio de Educacion Nacional.* Bogota, Colombia.

Fibe Frank, I. H. (1998). Generating Accurate Rules Sets Wiyhout Global Optimization. *Fifteenth International Conference on Machine Learning*, (págs. 144 -151). Madison, Wisconsin, USA.

Fontecha Ariza, C. (2007). *Análisis de los resultados de las pruebas SABER e I.C.F.E.S. en los componentes de matematica y lenguaje y su efecto en los estandares de calidad de la educacion en los colegios oficiales de las localidades de Usaquen y Ciudad Bolivar de bogota.* Bogota.

Gomez, ,. M., & Jaramillo, ,. J. (2017). *Descubrimiento de factores asociados al desempeño en las pruebas saber 5 con tecnicas descriptivas de mineria de datos.* Pasto, Nariño, Colombia.

Gonzales, C. J. (2012). Rendiiento academico y factores asociados. Aportaciones de algunas evaluaciones a gran escala. *Bordon 64(2) ISSN: 0210-5934.*

Gonzales, C., Caso, J., Diaz, K., & Lopez, M. (2012). *Renddimiento academico y factores asociados.*

ICFES. (2009). *Saber 5 y 9 Sintesis de resultados de Factores asociados.* Bogota, Colombia.

ICFES. (2013). *Alineacion del examen SABER 11. Sistema Nacional de Evaluacion Estandarizada de la Educacion.* Bogota. Colombia.

ICFES. (2014). *Alineacion del examen SABER 11. Lineamientos generales 2014-2 Sistema Nacional de Evaluacion Estandarizado de la Educacion.* Bogota, Colombia.

ICFES. (2018). *Documentacion de la prueba Saber 3, 5 y 9.* Bogota, Colombia.

ICFES. (2013). *Alineacion del examen SABER 11. Sistema Nacional de Evaluacion Estandarizada de la Educacion, Instituto Colombiano para la Evaluacion de la Educacion.* Bogota, Colombia.

ICFES. (2016). *Pruebas Saber 3. 5. 9. Sistema Nacional de Evaluacion Estandarizado de la Educacion en Colombia.* Bogota,Colombia.

Jimenez,, A., & Alvarez,, H. (2010). *Mineria de Datos en la Educacion. Universidad Carlos III de Madrid Avda.* Leganes, (Marid, España).

MEN. (28 de Enero de 2018). *Colombia aprende la red del conocimient*. Obtenido de Colombia aprende la red del conocimiento: http://www.colombiaaprende.edu.co/html/home/1592/article-89525.html

MEN. (18 de junio de 2019). *Colombia Aprende*. Obtenido de Centro Virtual de Noticias de Educacion.: https://www.mineducacion.gov.co/cvn/1665/w3-article-349894.html

MEN. (23 de junio de 2019). *Ministerio de Educacion*. Obtenido de Ministerio de Educacion: www.mineducacion.gov.co/1621/article-162392.html

Mr.Suhas G. Kulkarni, M. C. (2016). Big data analytics and e learning higher education. *International Journal on Cybernetics & Informatics (IJCI) Vol. 5, No. 1.*

Orea, S., Vargas, A., & Alonso, M. (2005). *Mineria de Datos: prediccion de la desercion escolar mediante el algoritmo de arboles de decicion y el algoritmo de los K vecinos mas cercanos. Ene, 779(73),33.*

Piñeros, ,. L., & Rodriguez, ,. (1999). School inputs in secondary education and their effects on academic achievement: a study in Colombia. *Human Development Department, World Bank.*

Rodriguez, H. (2007). El paradigma de las cpmpetencias hacia la educacion superior. *Facultad de Ciencias Economicas. Universidad Militar Nueva Granada. V. V. N. 1,145165*.

Rojas Rubio, ,. M. (1992). promocion automatica y fracaso escolar en colombia. *Red Academica No. 25 (Universidad Pedagogica).*

Science, D. (4 de septiembre de 2019). *Data Science de Novato a Experto*. Obtenido de http://datascience.esy.es/wiki/zeror/

Timaran, ,. R., Benavides, ,. J., & Hidalgo, ,. (octubre de 2018). *Discovering Factors Associated with Academic Performance of High School*. Obtenido de https://www.researchgate.net/publication/328769291: https://www.researchgate.net/publication/328769291

Timaran, R., Calderon, A., & Jimenez, j. (2013a). Aplicacion de Mineria de Datos en la extraccion de perfiles de desercion estudiantil. *Ventana Informatica, No. 28,2538*.

Timaran, R., Calderon, A., & Jimenez, J. (2013b). La Mineria de Datos como un metodo innovador para la deteccion de patrones de desercion en programas de pregrado en instituciones de Educacion Superior. *En Memorias Foro Mundial de Educacion en Ingenieria, WEEF2013 Cartagena, Colombia: ACOFI & IFEES.*

Timaran-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Arboles de desiciones para predecir factores asociados al desempeño academico de estudiantes de bachillerato en las pruebas saber 11. *Rev.investig.desarr.innov., 9(2)*,.

Torrado, M. (2000). *Educar para el desarrollo de las competencias. Una propuesta para feflexionar en competencias y proyecto pedagogico.*

Valero, S. (2009). *Aplicacion de tecnicas de Mineria de Datos para predecir desercion.* Obtenido de Unversidad Tecnologica de Izucar de Matamorros: http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf

Valero,, S., Salvador,, A., & Garcia,, M. (2010). *mineria de Datos: prediccion de la desercion escolar mediante el algoritmo de arboles de desicion y el algoritmo de los ka vecinos mas cercanos.* Obtenido de Unversidad Tecnologica de Izucar de Matamorros: http://www.utim.edu.mx/~svalero/docs/e1.pdf

Velasquez, W. (2013). *Estilos de aprendizaje y Rendimiento Academico.* Medellin, Colombia.

Villafañe, V. P. (2015). *Analisis del desempeño academico del examen de estado para el ingreso a la educacion superior aplicando mineria de datos.* Bogota, colombia.